TESE DE DOUTORAMENTO

POPULATION GENOMICS AS A TOOL FOR MANAGEMENT AND CONSERVATION OF BROWN TROUT (*SALMO TRUTTA*) IN THE IBERIAN PENINSULA

Adrián Casanova Chiclana

ESCOLA DE DOUTORAMENTO INTERNACIONAL DA UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

PROGRAMA DE DOUTORAMENTO EN MEDIO AMBIENTE E RECURSOS NATURAIS

LUGO

2021

DECLARACIÓN DO AUTOR/A DA TESE

D./Dna. **Adrián Casanova Chiclana**

Título da tese: **Population genomics as a tool for management and conservation of brown trout (Salmo trutta) in the Iberian Peninsula**

Presento a miña tese, seguindo o procedemento axeitado ao Regulamento, e declaro que:

1) A tese abarca os resultados da elaboración do meu traballo.
2) De ser o caso, na tese faise referencia ás colaboracións que tivo este traballo.
3) Confirmo que a tese non incorre en ningún tipo de plaxio doutros autores nin de traballos presentados por min para a obtención doutros títulos.
4) A tese é a versión definitiva presentada para a súa defensa e coincide a versión impresa coa presentada en formato electrónico

E comprométome a presentar o Compromiso Documental de Supervisión no caso de que o orixinal non estea na Escola.

En **Lugo**, **07 de Xuño de 2021**.

Sinatura electrónica

AUTORIZACIÓN DO DIRECTOR DA TESE

Population genomics as a tool for management and conservation of brown trout (*Salmo trutta*) in the Iberian Peninsula

D. Manuel Vera Rodríguez

D. Paulino Martínez Portela

INFORMA/N:

Que a presente tese, correspóndese co traballo realizado por D. Adrián Casanova Chiclana , baixo a miña dirección, e autorizo a súa presentación, considerando que reúne os requisitos esixidos no Regulamento de Estudos de Doutoramento da USC, e que como director desta non incorre nas causas de abstención establecidas na Lei 40/2015.

De acordo co indicado no Regulamento de Estudos de Doutoramento, declara tamén que a presente tese de doutoramento é idónea para ser defendida en base á modalidade de Monográfica con reproducción de publicaciones, nos que a participación do/a doutorando/a foi decisiva para a súa elaboración e as publicacións se axustan ao Plan de Investigación.

En Lugo, 7 de xuño de 2021

AUTORIZACIÓN DO TITOR DA TESE

Population genomics as a tool for management and conservation of brown trout (*Salmo trutta*) in the Iberian Peninsula

D. Jesús Ramón Aboal Viñas

INFORMA/N:

Que a presente tese, correspóndese co traballo realizado por D. Adrián Casanova Chiclana, baixo a miña titorización, e autorizo a súa presentación, considerando que reúne os requisitos esixidos no Regulamento de Estudos de Doutoramento da USC, e que como director desta non incorre nas causas de abstención establecidas na Lei 40/2015.

De acordo co indicado no Regulamento de Estudos de Doutoramento, declara tamén que a presente tese de doutoramento é idónea para ser defendida en base á modalidade de Monográfica con reproducción de publicaciones, nos que a participación do/a doutorando/a foi decisiva para a súa elaboración e as publicacións se axustan ao Plan de Investigación.

En Santiago de Compostela, 7 de xuño de 2021

Este traballo foi financiado grazas as axudas de apoio á etapa predoutoral da Consellería de Cultura, Educación e Ordenación Universitaria e a Consellería de Economía, Emprego e Industria a través da Axencia Galega de Innovación (GAIN), cofinanciadas parcialmente polo programa operativo FSE Galicia 2014-2020, na convocatoria do ano 2017 (Nº expediente: ED481A-2017/091).

ABSTRACT

Brown trout (*Salmo trutta*) is a cold-water salmonid with ecological, commercial and recreational importance. Previous genetic studies highlighted a notable genetic structure in natural populations, which is a fundamental biological feature for the conservation of the species. Nevertheless, many factors threaten this genetic richness. One of the most important is genetic introgression, that originated from the introduction of aquaculture individuals in the wild environment in the last decades. This genetic erosion can have adverse consequences in brown trout adaptation capacity, especially in the context of a reduction of available habitats driven, among other causes, by global warming.

This doctoral thesis had two main strands: (1) A bioinformatic benchmark to evaluate the biological conclusion robustness drawn from diverse SNP panels derived from different building-loci pipelines. For this strand, a broad of five aquatic species representing different genomic and/or population structure scenarios was used. This strand was performed in a context in which, due to the irruption of new techniques as RAD-seq (Restricion site-associated DNA sequencing) for the library preparation and Next Generation DNA Sequencing (NGS), the volume of data generated has grown exponentially in the last decade, promoting the development of new bioinformatics tools to process it. On the other side, (2) a genomic approach was used for the first time on brown trout populations from the Iberian Peninsula, to evaluate genetic diversity levels, population structure, natural hybridization patterns, evaluation of population introgression with aquaculture individuals of the same species and the detection of candidate genomic regions under selection pressure. The information obtained will allow the improvement of current management and conservation strategies of wild brown trout genetic resources.

KEYWORDS

Brown trout, bioinformatic benchmark, RAD-seq, Stacks 2, conservation genomics.

RESUMO

A troita común (*Salmo trutta*) é un salmónido de auga fría con importancia ecolóxica, comercial e recreativa. Estudos xenéticos anteriores puxeron de manifesto unha notable estrutura xenética nas poboacións naturais, o que constitúe unha característica biolóxica fundamental para a conservación da especie. Con todo, moitos factores ameazan esta riqueza xenética. Un dos máis importantes é a introgresión xenética, orixinada pola introdución de individuos de acuicultura no medio natural nas últimas décadas. Esta erosión xenética pode ter consecuencias adversas na capacidade de adaptación da troita común, especialmente no contexto dunha redución dos hábitats dispoñibles impulsada, entre outras causas, polo quecemento global.

Esta tese de doutoramento presentou dúas liñas principais: (1) Unha referencia bioinformática para avaliar a robustez das conclusións biolóxicas derivadas da análise de diversos paneis de SNPs procedentes de diferente *software* bioinformático para a construción de loci. Para esta liña, empregáronse cinco especies acuáticas que representaban diferentes escenarios xenómicos e/ou de estrutura poboacional. Esta investigación realizouse nun contexto no que, debido á irrupción de novas técnicas como RAD-seq (*Restricion site- associated DNA sequencing*) para a preparación de librarías e *Next Generation DNA Sequencing* (NGS), o volume de datos xerados creceu exponencialmente na última década, promovendo o desenvolvemento de novas ferramentas bioinformáticas para o seu procesamento. Doutra banda, (2) utilizouse por primeira vez unha aproximación xenómica en poboacións de troita común da Península Ibérica, para avaliar os niveis de diversidade xenética, estrutura poboacional, patróns de hibridación natural, avaliación da introgresión poboacional con individuos de acuicultura da mesma especie e a detección de rexións xenómicas candidatas a estar baixo presión selectiva. A información obtida permitirá mellorar as actuais estratexias de xestión e conservación dos recursos xenéticos da poboacións naturais de troita común.

PALABRAS CHAVE

Troita común, referencia bioinformática, RAD-seq, Stacks 2, xenómica da conservación.

RESUMEN

La trucha común *(Salmo trutta*) es un salmónido de agua fría con importancia ecológica, comercial y recreativa. Estudios genéticos anteriores pusieron de manifiesto una notable estructura genética en las poblaciones naturales, lo que constituye una característica biológica fundamental para la conservación de la especie. Sin embargo, muchos factores amenazan esta riqueza genética. Uno de los más importantes es la introgresión genética, originada por la introducción de individuos de acuicultura en el medio natural en las últimas décadas. Esta erosión genética puede tener consecuencias adversas en la capacidad de adaptación de la trucha común, especialmente en el contexto de una reducción de los hábitats disponibles impulsada, entre otras causas, por el calentamiento global.

Esta tesis doctoral presentó dos líneas principales: (1) Una referencia bioinformática para evaluar la robustez de las conclusiones biológicas derivadas del análisis de diversos paneles de SNPs procedentes de diferente *software* bioinformático para la construcción de loci. Para esta línea, se emplearon cinco especies acuáticas que representaban diferentes escenarios genómicos y/o de estructura poblacional. Esta investigación se realizó en un contexto en el que, debido a la irrupción de nuevas técnicas como RAD-seq (*Restricion site-associated DNA sequencing*) para la preparación de librerías y *Next Generation DNA Sequencing* (NGS), el volumen de datos generados ha crecido exponencialmente en la última década, promoviendo el desarrollo de nuevas herramientas bioinformáticas para su procesamiento. Por otro lado, (2) se ha utilizado por primera vez una aproximación genómica en poblaciones de trucha común de la Península Ibérica, para evaluar los niveles de diversidad genética, estructura poblacional, patrones de hibridación natural, evaluación de la introgresión poblacional con individuos de acuicultura de la misma especie y la detección de regiones genómicas candidatas bajo presión selectiva. La información obtenida permitirá mejorar las actuales estrategias de gestión y conservación de los recursos genéticos de las poblaciones naturales de trucha común.

PALABRAS CLAVE

Trucha común, referencia bioinformática, RAD-seq, Stacks 2, genómica de la conservación.

RESUMO AMPLO

1 INTRODUCCIÓN

Troita común (*Salmo trutta*)

A troita común (*Salmo trutta*) é unha especie de peixe de auga doce, pertencente ao orde Salmoniformes. Esta orde está formada por unha única familia, Salmonidae dentro da cal existen dúas especies autóctonas en España, pertencentes ao xénero *Salmo*: a troita común (*Salmo trutta*) e o salmón atlántico (*Salmo salar*) (Doadrio 2001). Varias especies do xénero *Salmo* son filoxenéticamente próximas polo que adoita considerarse a *S. trutta* coma un complexo multiespecie (Gratton *et al.* 2014; Splendiani *et al.* 2019; Vera *et al.* 2011).

Esta especie vive en regatos rápidos, ríos e lagos fríos e ben osixenados (7 mg O_2/L e unha saturación de osíxeno do 80%; Cobo *et al.* 2015; Eklöv *et al.* 1999). O seu rango de distribución natural abrangue principalmente Eurasia (Freyhof 2011), desde o norte de Noruega e o nordeste de Rusia oriental (Bernatchez *et al.* 1992) ata o Cordal do Atlas no norte de África (Sanz 2018). Debido ás introducións mediadas polo ser humano, a troita común ten un área de distribución foránea que inclúe o extremo oriental de Asia, Oceanía, África e América (Casal 2006).

A troita común presenta unha ampla variedade fenotípica relacionada coa diversidade do medio no cal vive e o seu comportamento migratorio. Seguindo criterios ecolóxicos e de ciclo vital identificáronse tradicionalmente tres formas de troita común: (1) residente, (2) anádroma (reo) e (3) lacustre. Os reos pódense atopar na vertente atlántica, non así na mediterránea, sendo o río Limia (42º N) o seu límite meridional de ocorrencia (Bouza *et al.* 1999). Estes exemplares aliméntanse principalmente preto da costa, non moi lonxe da desembocadura dos ríos de procedencia (Kottelat e Freyhof 2007). A troita é un peixe de tamaño medio que en España pode acadar os 60 cm e os 10 kg, acadando os individuos anádromos e lacustres maiores tamaños (45-60 cm de lonxitude estándar, SL), que os residentes (20-30 cm SL, Kottelat y Freyhof 2007). A troita común pode presentar manchas escuras no dorso e nos laterais do corpo, as veces rodeadas por un halo pálido. Os reos tralo esguinado

presentan unha coloración prateada e as manchas son menos visibles que na forma residente (Schultz 2003). Tralo esguinado as troitas migran ao mar para alimentarse e retornan aos ríos para desovar. Existen poboacións de troita común formadas por individuos migrantes e non migrantes vivindo en simpatría (Jonsson e Jonsson 2011) podendo reproducirse entre ambos morfotipos (Ruzzante *et al.* 2001). Este fenómeno de coexistencia é coñecido como migración parcial (Lundberg 1988). Cabe destacar que hai migracións dentro das diferentes formacións de auga doce, i.e. potamodromía (Jonsson e Jonsson 2011) ou diferentes ciclos de vida posibles (ver Fig. 1 en Ferguson *et al.* 2019).

Conservación da especie

A troita común é considerada unha especie bandeira (Denic e Geist 2010) e/ou paraugas (Lobón-Cerviá 2018). O emprego destes termos para referirse á troita común reflicte a importancia para o funcionamento dos ecosistemas fluviais e para o ser humano. Esta especie é un depredador de nivel superior nos ecosistemas fluviais nos que habita (Sánchez-Hernández 2016), presentando ademais unha dieta moi ampla (especie eurifáxica; Alonso *et al.* 2017). Habitualmente, o tamaño das presas aumenta co tamaño do individuo debido ao aumento en lonxitude do peixe e a unha maior apertura bucal (Sánchez-Hernández 2016). Cabe destacar ademais o papel da troita común como hóspede co mexillón de río (*Margaritifera margaritifera*), unha especie en perigo de extinción (Moorkens *et al.* 2017) que resulta ser unha compoñente integral dos ecosistemas fluviais xa que filtran partículas e serven de sustrato/hábitat para algas e invertebrados bentónicos, nunha Península Ibérica con baixa riqueza específica de mexillóns de auga doce (ver Fig. 7 en Lopes-Lima *et al.* 2016). Este bivalvo presenta nunha fase do seu ciclo vital unha larva especializada (i.e. gloquidio) que normalmente debe parasitar a un peixe hóspede sobre o cal enquístase e realiza a metamorfose, sendo estes hóspedes específicos *S. trutta* e *S. salar*. Desde unha perspectiva humana a troita común presenta unha elevada importancia debido a que é unha especie emblemática para a pesca recreativa (FAO 2012), a cal ten un apreciable peso en España como mostra o case medio millón de licenzas de pesca recreativa expedidas no 2018 (Ministerio para la Transición Ecológica y el Reto Demográfico, MITERD 2020). O impacto económico da pesca recreativa

pode ser maior cando a práctica da mesma implica desprazamentos alén da rexión de orixe dos pescadores, vinculado isto co turismo activo, ecoturismo e turismo rural (FAO 2012) podendo presentar un importante impacto económico nalgunhas rexións (por exemplo, Escocia, Butler *et al.* 2009; Gotland, Blicharska e Rönnbäck 2018). Para facer fronte á demanda dos pescadores cada ano realízanse soltas de exemplares ou ovas de troita común, representando o 60% do total. Non obstante, a troita común, alén do seu rango de distribución nativo é causa de diferentes inconvenientes coa biodiversidade nativa, sendo considerada polo Grupo de Especialistas en Especies Invasoras (ISSG) como unha das 100 peores especies exóticas invasoras do mundo.

O estado de conservación da troita común segundo a IUCN é de "Preocupación menor" (Freyhof 2011), mais presenta unha serie de ameazas relacionadas coa fragmentación do hábitat, contaminación acuática, quecemento global, sobrepesca e introgresión debido a soltas de piscifactoría. Unha das características dos ecosistemas fluviais e a conectividade lonxitudinal que pode verse interrompida polas construción de diferentes infraestruturas entre as que destacarían as presas que varían as súas características acorde os seus fins (e.g. obtención de enerxía hidroeléctrica, regulación do nivel da auga), afectando aos movementos desta especie, illando poboacións e reducindo a presencia de reos (Birnie-Gauvin *et al.* 2017; Heggenes e Røed 2006). No tocante á contaminación acuática a troita común considérase unha especie sensible (ver Luckenbach *et al.* 2001; Pickering 1989) se ben, reportáronse algúns casos de tolerancia en ecosistemas altamente contaminados (ver Durrant *et al.* 2011). Un aumento de temperaturas asociado ao quecemento global suporía a perda de hábitats adecuados para a troita común, especialmente nos límites meridionais do seu rango de distribución (Almodóvar *et al.* 2012). Finalmente, a repoboación pode poñer en perigo ás poboacións nativas ao alterar as adaptación locais ao introducir a selección causada pola domesticación (Pinter *et al.* 2019). Na actualidade as tendencias actuais son a acuicultura da conservación (e.g. mellorar as poboacións silvestres con individuos de maior similitude xenética) ou a liberación de individuos triploides para a pesca (i.e. individuos estériles). Non obstante a triploidía non ten un éxito do 100% e son necesarios métodos

eficaces para detectar individuos diploides para evitar a liberación de miles de individuos fértiles procedentes de piscifactorías (Sanz *et al.* 2020).

Xenética e xenómica da troita común.

A troita común é unha especie cunha alta diferenciación xenética entre poboacións, incluso entre as pertencentes a tramos próximos dun mesmo río (Bouza *et al.* 1999; Fernández-Cebrián *et al.* 2014), sendo unha das especies de vertebrados máis estruturadas ($F_{ST} > 0,60$; Ferguson 1989). Con todo, esta regra xeral non sempre se cumpre (Heggenes e Røed 2006). Na Península Ibérica observouse esta alta diferenciación entre concas fluviais da vertente mediterránea e atlántica utilizando diferentes marcadores moleculares (e.g. alozimas, Bouza *et al.* 1999; microsatélites, Martínez *et al.* 2007).

Esta especie presenta unha filoxeografía complexa, complicada por procesos de introgresión entre liñaxes inducidas polo ser humano (Sanz 2018). Inicialmente identificáronse cinco liñaxes mitocondriais principais empregando a rexión de control do ADN mitocondrial (CR mtDNA; Bernatchez *et al.* 1992): Adriático (AD), Atlántico (AT), Danubio (DA), Marmoratus (MA) e Mediterráneo (ME). Desde entón, varios estudos reportaron liñaxes adicionais restrinxidas á Península Ibérica no caso da liñaxe do Duero (DU; Vera *et al.* 2010); a liñaxe do Tigris en Turquía (TI; Sušnik *et al.* 2005); e unha oitava liñaxe norteafricana (Tougard *et al.* 2018), anteriormente considerado unha especie diferente (i.e. *Salmo macrostigma*). A rexión de control do ADNmt foi o principal marcador molecular utilizado para identificar conxuntos filoxenéticos xeograficamente estruturados e os haplogrupos de ADNmt empregáronse como aproximacións de unidades evolutivas significativas (ESUs) combinados coa información achegada por outros marcadores moleculares para o estudo do impacto das repoboacións, a diversidade xenética das poboacións silvestres, estrutura poboacional, etcétera (e.g. microsatélites, Vera *et al.* 2013; polimorfismos de nucleótido único, SNPs; Pustovrh *et al.* 2011).

Dentro dos peixes, a hibridación vese facilitada por características reprodutivas como a fecundación externa e as zonas de desova comúns (Hubbs 1955). Ao tratar os fenómenos de hibridación e introgresión na troita común, caben destacar tres escenarios diferentes: (1) a introgresión xenética das

poboacións silvestres coas procedentes de soltas de piscifactoría; (2) a hibridación natural entre liñaxes evolutivas (contactos secundarios); e (3) a hibridación con especies estreitamente relacionadas (e.g. *S. salar*). A liberación de troitas de piscifactoría no medio natural foi unha práctica de xestión aplicada desde principios do século XX. As repoboacións nas concas fluviais europeas realizáronse principalmente cun stock de orixe centroeuropeo (Martínez *et al.* 1993; Morán *et al.* 1991; Vera *et al.* 2018). Esta práctica realizouse tamén con outras especies de salmónidos (e.g. *Salvelinus fontinalis*, Lehnert *et al.* 2020; *S. salar*, Saltveit 2006). O segundo escenario mencionado está relacionado coa hibridación natural entre liñaxes e/ou subespecies, aínda que a validez da taxonomía de *Salmo trutta* en parte é controvertida. Na Península Ibérica foron descritos dous escenarios principais de hibridación natural: (1) entre as liñaxes AT e DU na conca do Duero e do Miño-Sil cun patrón de distribución parapátrica (Martínez *et al.* 2007; Vera *et al.* 2010; Vilas *et al.* 2010); e (2) entre AD e ME cun patrón de distribución en mosaico (Sanz Ball-Llosera *et al.* 2002). O último escenario exposto incluiría: (1) a hibridación interespecífica natural, e (2) a hibridación interespecífica inducida pola actividade humana. No primeiro, probablemente o caso máis relevante sería a hibridación entre *S. trutta* e *S. salar* no norte de España onde ambas especies viven en simpatría (García de Leaniz e Vespoor 1989). A segunda sería cando dúas especies de salmónidos non vivían orixinalmente en simpatría, sendo unha delas especie non autóctona (Cucherousset *et al.* 2008).

A irrupción da secuenciación de próxima xeración (NGS) e da secuenciación de ADN asociada a sitios de restrición (RAD-seq) na última década levou á identificación de miles de marcadores moleculares (i.e. SNPs) e ao seu xenotipado conducindo a unha transición desde un enfoque xenético a outro xenómico. Estas novas técnicas supuxeron unha revolución para especies non modelo, que non presentaban un xenoma de referencia, representando estas a maioría dos casos. O volume de datos implicou a súa vez un desenvolvemento de novos modelos e programas bioinformáticos para o seu procesamento.

Ata a data, realizáronse diferentes estudos seguindo estratexias xenómicas na troita común, algúns deles con chips de SNPs (ver Bekkevold

et al. 2019; Linløkken *et al.* 2017). Os diversos aspectos abordados están relacionados coa avaliación da estrutura xenética e a identificación das unidades de conservación, a avaliación e o seguimento da repoboación, a detección da variación adaptativa (i.e. sinaturas de selección) nunha ampla gama de características (por exemplo, o crecemento individual), entre outros. Ata a data non se abordou ningunha investigación que siga un enfoque de xenómica poboacional con poboacións de troita común da Península Ibérica. Ata o 2015, a principal forma de construír *loci* e xenotipado con ferramentas bioinformáticas na troita común utilizando RAD-seq era empregando un enfoque *de novo* (i.e. sen xenoma de referencia). O xenoma do salmón atlántico, a especie conxénere, está dispoñible desde 2015 grazas á iniciativa do Proxecto de Cooperación Internacional para a Secuencia do Xenoma do Salmón Atlántico (ICSASG). Realizáronse aproximacións ao xenoma de referencia nos datos RAD-seq da troita común utilizando o salmón atlántico como xenoma de referencia (véxase Paris *et al.* 2017). Aínda que os xenomas de referencia de especies estreitamente relacionadas poden utilizarse para obter SNPs que permitan estimar diferentes parámetros de xenética de poboacións, unha maior consistencia require empregar un xenoma de referencia propio da especie. Por exemplo, os xenomas propios son recomendables para obter a maior resolución posible (é dicir, o número e a posición dos SNPs) para estudar os patróns de hibridación ao longo do xenoma. En xuño de 2019 estivo dispoñible no *National Center for Biotechnology Information* (NCBI) o primeiro xenoma da troita común. Este logro produciuse no marco da iniciativa "25 xenomas para 25 anos" e foi promovido polo Instituto *Wellcome Sanger* (https://www.sanger.ac.uk/collaboration/25-genomes-25-years/). Dado que está dispoñible desde hai pouco máis dun ano, hai moi poucas investigacións publicadas que utilicen o xenoma de referencia de troita común.

2 OBXECTIVOS

A presente tese tivo como obxectivo global estudar a estrutura xenética da troita común (*Salmo trutta*) nas vertentes atlántica e mediterránea da Península Ibérica mediante un enfoque de xenómica de poboacións. Estudáronse os patróns de hibridación natural e introgresión mediada pola

actividade humana e considerouse toda a información obtida para a xestión e conservación dos recursos xenéticos autóctonos desta especie na Península Ibérica. Os obxectivos específicos foron:

1. **Avaliación dos diferentes programas bioinformáticos para a identificación e xenotipado de SNPs para acadar interpretacións biolóxicas consistentes a traveso dun enfoque de xenómica poboacional.**

Para a acadar este obxectivo realizouse unha comparación entre os resultados de dous ferramentas bioinformáticas para a construción de loci (*building-loci pipelines*), *Stacks 2* (Catchen *et al.* 2011, 2013; Rochette *et al.* 2019) e *Meyer's 2b-RAD v2.1 pipeline (*Wang *et al.* 2012*)*. A primeira é a máis empregada neste tipo de investigacións e a segunda é a orixinal para traballar con 2bRAD-seq (Wang *et al*. 2012). Empregáronse diferentes especies acuáticas que representaban diferentes escenarios xenómicos e/ou de estruturas poboacionais. Empregáronse dúas especies de bivalvos, berberecho (*Cerastoderma edule)* e ameixa xaponesa (*Ruditapes philippinarum*) e tres especies de peixes: troita común (*Salmo trutta*), bagre (*Rhamdia quelen*) e melgacho (*Scyliorhinus canicula*). Avaliáronse catro paneis de SNPs en cada especie para probar tanto os programas bioinformáticos empregados para a construción dos loci e o seu xenotipado como os criterios de selección de SNPs (i.e. filtrado empregado). Ademais, empregouse xenoma de referencia como control nas especies nas cales había un dispoñible (i.e. ameixa xaponesa e troita común).

Os resultados obtidos nos diferentes casos de estudo analizados mostraron que as *building-loci pipelines* empregadas non teñen unha influencia substancial na estimación dos parámetros poboacionais e as súas interpretacións biolóxicas derivadas. As pequenas diferencias observadas entre algúns paneis *de novo* e os derivados do emprego dun xenoma de referencia poderían resolverse mellorando os pasos de filtrado de SNPs. De todos os xeitos, estes resultados non deberían xeneralizarse e as persoas usuarias deberían contextualizar os *building-loci pipelines* aos seus escenarios xenómicos. Unha recomendación sería probar os parámetros dos *building-loci pipelines* e os pasos de filtrado de SNPs cun subconxunto das mostras para

aforrar recursos de soporte físico e tempo de computación. A mellor configuración de parámetros sería aquela que conducise aos máis consistentes a traveso das diferentes réplicas e empregando os parámetros poboacionais que correspondan aos obxectivos de investigación. A pesares de implicar un consumo de tempo, este paso preliminar melloraría a solidez dos resultados e as conclusións biolóxicas derivadas ao mellorar o uso das ferramentas bioinformáticas.

2. Análise da introgresión xenómica das poboacións silvestres por parte das poboacións de piscifactoría nos ecosistemas inestables da vertente mediterránea.

3. Estudo preliminar da hibridación natural nun contacto secundario na vertente atlántica.

4. Mellora das estratexias de conservación e xestión para manter a integridade dos liñaxes autóctonos e recuperar as poboacións naturais en zonas con alta taxa de introgresión.

Para acadar estes tres obxectivos diferentes, mostras de troita común procedentes da Península Ibérica foron tomadas tanto da vertente atlántica (concas do ríos Miño e Duero) coma da vertente mediterránea (rexión Pirenaica), ademais de troitas procedentes dun centro piscícola (Bagà). Nesta mostraxe atopábanse representadas as diferentes liñaxes mitocondriais da Península Ibérica (AD, AT, DU, ME), poboacións naturais afectadas con diferente intensidade por soltas desde piscifactorías e zonas de hibridación natural. Empregouse *Bowtie 1* (Langmead 2009) e *Stacks 2* xunto co xenoma de referencia de troita común para a obtención de loci e a detección e xenotipado dos SNPs. Trala obtención dun panel robusto con miles de SNPs realizouse unha serie de análises bioinformáticas para determinar o grao de afección por soltas de piscifactoría desde un enfoque xenómico, introgresión ao longo do xenoma, diversidade xenética, estrutura poboacional, estimación do censo efectivo, detección de pegadas de selección e anotación funcional.

Deseñouse unha ferramenta molecular para avaliar os efectos da repoboación nas poboacións naturais de troita común, permitindo detectar a introgresión a nivel individual e poboacional dun xeito máis barato e cunha

maior resolución, grazas ao elevado incremento de marcadores moleculares empregados, respecto á metodoloxía empregada ata o momento coas poboacións da Península Ibérica (emprego do locus diagnóstico *LDH-C**). Coa resolución xenómica acadada puidéronse detectar tamén pequenas rexións de introgresión ao longo do xenoma, inferindo a ancestralidade local a nivel cromosómico.

Todas as poboacións afectadas por repoboación estiveron en equilibrio Hardy-Weinberg, o que suxire que as variantes alélicas dos stocks de piscifactoría introducidas durante as últimas décadas integráronse e aleatorizáronse no patrimonio xenómico das poboacións silvestres.

Detectáronse grandes diferencias nos valores de diversidade xenética entre as diferentes poboacións de troita común na Península Ibérica. Confirmouse o patrón de aumento da diversidade xenética seguindo un gradiente leste-oeste e sur-norte, cos valores máis altos na conca hidrográfica do Miño. Este patrón estaría en consonancia coas rexións xeográficas cunhas condicións ambientais máis adecuadas para esta especie.

Detectouse unha fonda estruturación entre as poboacións pertencentes a diferentes concas, incluso entre poboacións próximas. A maior diferenciación xenética detectouse entre as vertentes atlántica e mediterránea, consecuencia da ausencia de fluxo xénico debida ao illamento xeográfico existente. Dentro da vertente atlántica, as concas do Duero e do Miño apenas terían fluxo xénico debido á escaseza de reos que poidan comunicar ambas concas, de habelos. Empregando diferentes subconxuntos de SNPs nucleares, non se detectaron as zonas de hibridación natural previamente reportadas en estudos anteriores. Isto suxeriría que só quedan rastros dos eventos de hibridación por contacto secundario debido á forte deriva xenética asociada ao pequeno censo efectivo nas poboacións de troita común.

Detectáronse SNPs que poden estar baixo presión selectiva (i.e. *outliers*) empregando diferentes niveis xerárquicos, na súa maioría de selección estabilizadora entre as vertentes mediterránea e atlántica, namentres que baixo selección diverxente na vertente atlántica tanto dentro como entre concas (i.e. Miño-Duero). De todos xeitos, a detección bioinformática destes *outliers* sería o paso previo a análises máis exhaustivas para a validación de presión

selectiva nesas rexións xenómicas, mediante análises de expresión xénica en experimentos de ambiente común (*common-garden*).

A modo de peche, a información obtida da análise xenómica poboacional realizada nesta tese será de gran valor para deseñar as directrices de conservación das poboacións de troita común da Península Ibérica. Por exemplo, poderíanse aplicar diferentes políticas de conservación en función do grao de ancestralidade cos *stocks* de piscifactoría nas poboacións naturais de troita común: (1) erradicación das poboacións naturalizadas das soltas de piscifactoría que aínda poidan existir; e (2) establecemento de diferentes cotas de pesca (individuos por persoa e día) en función dos diferentes graos de introgresión detectados, reducindo as cotas nos ríos con poboacións naturais non afectadas xenéticamente polas soltas de piscifactoría. A detección de zonas de hibridación natural, non inducidas directamente polas actividades humanas, sería interesante desde o punto de vista da conservación, para protexer os procesos naturais que poidan aumentar a diversidade xenética e a viabilidade das poboacións. Algunhas das poboacións analizadas mostraron un baixo censo efectivo, o que podería levar á definición das poboacións de troita común como vulnerables segundo este criterio e posteriormente ao establecemento de medidas de conservación como moratorias temporais de pesca ou redución de cotas en unidades hidrográficas ben establecidas (e.g. arroios, tramos fluviais). A alta diferenciación xenética detectada nas poboacións de troita común pertencentes a unha mesma conca hidrográfica, implicaría que a repoboación baixo principios de conservación (utilizar individuos xenéticamente similares ás poboacións repoboadas) contemplada nos plans e leis de conservación pode ser difícil de implementar dunha forma estrita. En caso de aplicarse, deberían utilizarse poboacións non afectadas por eventos de introgresión, segundo os enfoques de ascendencia global e local. A definición de refuxios de troita común, nos que debería prohibirse a repoboación ou a pesca, en tramos de diferentes ríos con boas condicións ecolóxicas podería ser unha acción complementaria para a conservación.

De todos os xeitos, para adoptar medidas de xestión e conservación, sería necesario traballar con información actualizada sobre unha serie de poboacións representativas de diferentes unidades da Península Ibérica (e.g. ríos, concas hidrográficas, Unidades de Xestión). O ideal sería que as

estratexias fosen consensuadas e compartidas entre as diferentes Administracións para un plan de conservación integrado máis aló dos límites administrativos. A información xenómica obtida, combinada cos resultados doutras disciplinas (e.g. ecoloxía), debería axudar a tomar as mellores decisións de xestión. En calquera caso, ningunha medida de conservación debería abordarse sen os contextos ecolóxicos (e.g. o estado de conservación dos hábitats) e socioeconómicos asociados (e.g. a lexislación vixente e os usos económicos e culturais existentes).

Relación de artigos que aportan contido á presente Tese de doutoramento. Os contidos corresponden ao capítulo 3 e capítulo 5.1 da presente Tese de doutoramento.

Título:	Low impact of different SNP panels from two building-loci pipelines on RAD-Seq population genomic metrics: case study on five diverse aquatic species
Ano:	2021
Revista:	BMC Genomics
Volumen, Páx.:	22, Article number 150
DOI:	https://doi.org/10.1186/s12864-021-07465-w
Contribución do doutorando:	Adrián Casanova Chiclana codeseñou o estudo, realizou a meirande parte das análises bioinformáticas e escribiu a versión inicial do manuscrito ademais das diferentes versións coa contribución do resto de coautores
Índices/s de calidade:	4.093 5-year Citation Impact Factor

AGRADECEMENTOS

En primeiro lugar gustaríame dar as grazas aos meus directores de tese Manuel Vera e Paulino Martínez. Por toda a súa axuda, dedicación e formación nesta etapa. Todas as palabras de agradecemento para vós serían escasas.

Para todas as persoas coas que tiven o privilexio de formar equipo neste período no Departamento de Xenética. Profesionais das que puiden aprender e persoas coas que puiden compartir.

A Hurzos, por aquela chamada.

A todas as miñas amizades, a miña benquerida familia, A Asociación Cultural Náufragos do Paradiso e a todas as súas integrantes; por compartir a vida.

A Tamara, por ser unha estrela fugaz feita realidade.

E por suposto, e con todo o agarimo que pode conter un fillo. Mil grazas a miña mai, María. Por todo o bo que teño.

This page is intentionally left blank.

No tracks in the road, nothing living anywhere. The fire blackened boulders like the shapes of bears on the starkly wooded slopes. He stood on a stone bridge where the waters slurried into a pool and turned slowly in a grey foam. Where once he had watched trout swaying in the current, tracking their perfect shadows on the stones beneath.

The Road (2006). Cormac McCarthy.

A vida resúmese en saltar presas.
Troita común.

INDEX

1 GENERAL INTRODUCTION

1.1 BROWN TROUT

1.1.1 Taxonomy and distribution

Brown trout (*Salmo trutta*) is a freshwater fish, belonging to the order Salmoniformes. This order is constituted by a single family, Salmonidae (Fig. 1). Within Salmonidae, there are two native species in Spain belonging to the genus *Salmo*: brown trout (*Salmo trutta*) and Atlantic salmon (*Salmo salar*) (Doadrio 2001). This genus, already established by Linnaeus, would be constituted by about 40 species, some of them recently discovered such as *Salmo kottelati* (Turan *et al.* 2014), or others like *Salmo visovacensis*, whose validity is questionable, and some authors suggest its inclusion in another species, *Salmo farioides* (Bianco 2014).

Since many species of the genus *Salmo* are phylogenetically close, *S. trutta* is usually considered as a multi-species complex (Gratton *et al.* 2014; Splendiani *et al.* 2019; Vera *et al.* 2011). The stricter riverine habitat and higher degree of geographic isolation of brown trout populations than in Atlantic salmon, is responsible for the higher structure and phenotypic variation of brown trout. This has caused an important taxonomic controversy with this species (Jonsson and Jonsson 2011), especially when taxonomy was based only on phenotypes.

Brown trout lives in cold and well oxygenated swift streams, brooks rivers and lakes. Sea trout forage mostly close to the coast, not far from the mouth of natal rivers (Kottelat and Freyhof 2007). Its natural range covers mainly Eurasia (Freyhof 2011), from northern Norway and the north-eastern part of East Russia (Bernatchez *et al.* 1992) and to the Atlas Mountains in North Africa (Sanz 2018). Due to human-mediated introductions, brown trout has an alien range including the east end of Asia, Oceania, Africa, and America. These introductions have been registered in different databases (Casal 2006), such as the Database on Introductions of Aquatic Species (DIAS) dependent on Food and Agricultural Organization of the United Nations (FAO). The main motivation for brown trout introductions according to Welcomme (1988) would be for angling uses.

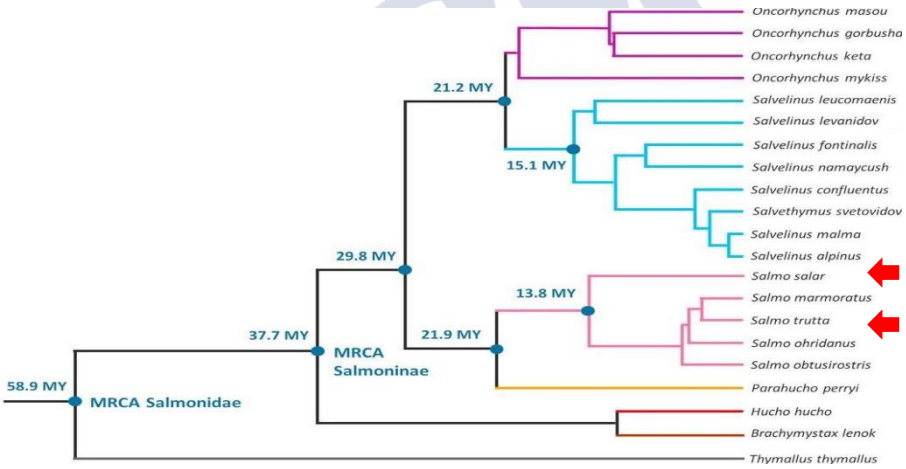

Figure 1. Phylogeny of the family Salmonidae. The common ancestor to *S. trutta* and *S. salar* (see red arrows) would be about 13 million years ago, during Miocene. Figure from Lecaudey *et al.* (2018).

1.1.2 Morphology

Brown trout has a wide phenotypic variety related to the diverse environment which they live and migratory behaviour. Following ecological or life-cycle criteria, three brown trout forms have been

traditionally identified: (1) resident, (2) anadromous, and (3) lacustrine. Brown trout is a medium size fish that in Spain can reach 60 cm and 10 kg (Doadrio 2001). Anadromous and lacustrine individuals can reach larger sizes (45-60 cm standard length, SL) than resident ones (about 20-30 cm SL; Kottelat and Freyhof 2007). Brown trout shows a brown, golden brown or olive tonality of its body. The belly can be yellowish or white (Schultz 2003). It can present a lot of dark spots on the back and sides of the body (Fig. 2), sometimes encircled by a pale halo (Fig. 3).

Figure 2. Brown trout specimen from Spain. Image taken by I. Doadrio (Doadrio 2001).

The brown trout coloration comes from the pigments existing in the different types of chromatophore cells: melanophores (black), xanthophores (yellow), erythrophore (red), etc (Djurdjevič *et al.* 2015).

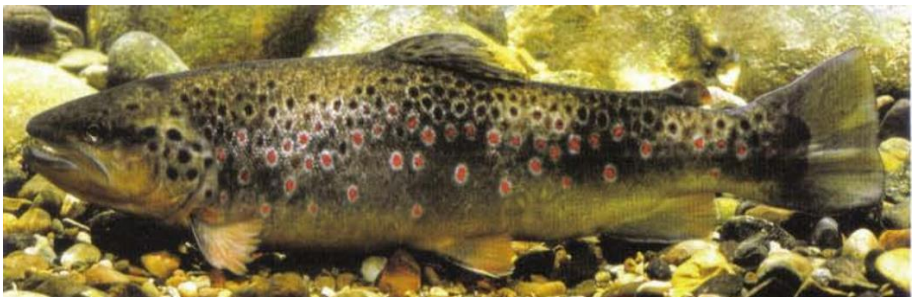

Figure 3. Brown trout specimen from Danube River. Image taken by A. Hartl (Kottelat and Freyhof 2007).

Anadromous brown trout has a silvery coloration (Fig. 4) and the spotting is less visible than in the resident form (Schultz 2003). This colour pattern would be related to the number and activity of iridophores during smoltification of sea trout (Cobo *et al.* 2015).

Figure 4. Brown trout specimen from lower Rhine, anadromous form after smoltification. Image taken by I. Steinmann (Kottelat and Freyhof 2007).

The morphology is variable, in accordance with a multi-morphic species. The head is big, maxilla reaches the posterior margin of the eye, and the posterior vomer is toothed in adults (Kottelat and Freyhof 2007). The number of gill rakers can vary within a range of 11-20 (Kottelat and Freyhof 2007; Langeland and Nøst 1995) among individuals and locations (Karakousis *et al.* 1991). The gills are protected by a bone structure, the operculum. The trunk ranges from operculum posterior edge to anus. In the trunk there are pectoral, pelvic and dorsal fins, the last one with 9-11 branched rays (Doadrio 2001). The tail ranges from the anus to the posterior caudal fin. In the tail there are one adipose, anal, and caudal fins. The caudal fin is homocercal in salmonids and is connected to the trunk by means of the tail peduncle. The lateral line, a sensorial organ, flanks the body from the tail peduncle to head.

1.1.3 Anadromy of Iberian brown trout

Two main brown trout morphotypes have been described living in Western Atlantic rivers: the anadromous (sea trout) and the resident

(Caballero *et al.* 2013). After smoltification, sea trout migrates to sea for feeding and return to freshwater for spawning. There may be populations consisting of migrant and non-migrant individuals, living in sympatry (Jonsson and Jonsson 2011) and being able to reproduce between them (Ruzzante *et al.* 2001). This phenomenon is known as partial migration (Lundberg 1988). It is worth noting that there are migrations in freshwater as well, denoted as potamodromy (Jonsson and Jonsson 2011), different life-histories are possible (see Fig. 1 in Ferguson *et al.* 2019).

Nowadays, the Iberian Peninsula is the southern distribution limit of sea trout (Bouza *et al.* 1999; Marco-Rius *et al.* 2013), this form being roughly absent southwards the 42º N parallel (Hamilton *et al.* 1989). The Limia River Basin is the southernmost limit of anadromy (SLA; Antunes *et al.* 2006), even though this limit was suggested to be shifted southwards to the Mondego River (see Caballero *et al.* 2018). Resident morphotypes would be the only existing south the SLA, where higher temperatures restrict populations to upstream river regions reducing the migratory potential (García-Marín and Pla 1996).

According to current knowledge, migration behaviour can happen depending on certain ecological and genetic factors (Caballero *et al.* 2013). Furthermore, epigenetic regulation during smoltification has been suggested (Morán *et al.* 2013) and some consistent outlier SNPs were detected to be associated with resident and migratory populations. This has suggested that migration in brown trout could be influenced by a set of candidate genes that appear to be shared with other salmonids (see Table 2 in Lemopoulos *et al.* 2018). It should be noted that sea trout could contribute to gene flow between river drainages (Bouza *et al.* 1999; Cobo *et al.* 2015).

1.1.4 Biology and ecology

Brown trout is a sexual dimorphic species with differences in secondary sexual characters (Reyes-Gavilán *et al.* 1997), but even in habitat use (Klemetsen *et al.* 2003). Brown trout normally reach sexual maturity between two and three years of life, later in females than in males (Alp *et al.* 2003). The reproduction is in autumn or winter and is phenotypically detectable by an increase in the gonadosomatic index (GSI; Jamalzadeh *et al.* 2013). The reproduction period is sooner at higher altitudes or latitudes due to the lower temperature of the water involving longer incubation periods (Cobo *et al.* 2015). The development time is usually measured in degree-days that accumulates the number of ºC through days of development (Ojanguren and Braña 2003) and that would be around 450 degree-days on average (Vøllestad and Lillehammer 2000). The incubation of eggs takes more than 40 days with normal freshwater temperatures in Iberian Peninsula (Doadrio 2001). Brown trout is highly selective in their choice of the spawning area and the gravel granulometry is an important factor (Barlaup *et al.* 2008). Females select the most appropriate spawning sites and dig the redd where the eggs are deposited (Kottelat and Freyhof 2007), generally in well-oxygenated waters because hatching of eggs requires a constant flow of water through the gravels to oxygenate them and remove waste products. Fertilization is external and occurs immediately after the eggs and sperm are expelled into the aquatic environment. In the Northern Hemisphere, the main period of growth is from April to November, when water temperatures are mostly between 5 and 20 ºC. In this range, salmonids feed enough, not only to maintain themselves but also for fattening and maturation (Cobo *et al.* 2015).

In fish, standard metabolic rate (SMR) is primarily limited by the capacity of the gills to extract oxygen from water. An increasing in oxygen uptake is necessary to do activities such as feeding, swimming,

etc (Eddy and Handy 2012). Like other fish, brown trout present skin breathing whose extent depends on environmental conditions (e.g. water temperature), but usually lower than 10% (Hill *et al.* 2016). Brown trout is susceptible to human activities that may affect water quality (e.g. pollution) and requires well oxygenated freshwater with higher concentrations than 7 mg O_2/L and oxygen saturation around 80% (Cobo *et al.* 2015; Eklöv *et al.* 1999). Lower concentrations can lead to delayed fry development and growth (Dumas *et al.* 2007) up to massive fish deaths. This species was evaluated as possible biomonitor in different studies (Lamas *et al.* 2007; Linde *et al.* 1998; Schmidt *et al.* 1999).

Brown trout has been considered a flagship (Denic and Geist 2010) or umbrella species (Lobón-Cerviá 2018). Although these terms have been used interchangeably in different studies (Caro 2010), they have different meanings. A flagship species would be an emblematic species selected to act as icon or symbol for a defined habitat, environmental campaign, etc; while an umbrella species could be species whose conservation confers protection to many naturally co-occurring species (Roberge and Angelstam 2004). The use of these terms to refer to the brown trout shows its importance for the functioning of the ecosystem and the value for humans.

One of the main ecological relationship would be predator-prey relationship. Brown trout is a top fish predator wherever it lives (Sánchez-Hernández 2016). Furthermore, it is an organism with a very broad diet (i.e euriphagic species; Alonso *et al.* 2017). The composition of brown trout diet can vary among individuals, populations, age classes, habitats, and seasons. Habitually, the size of the prey increases with the size of the individual due to the increase of fish length and mouth gape. This implies that dietary changes during the ontogenetic development of top fish predators involve changes in structural properties of food webs (Sánchez-Hernández 2016).

Beyond its trophic relationship as an upper-level predator in ecosystem it should be noted the condition of brown trout as a host (i.e parasitism relationship) with the freshwater pearl mussel (*Margaritifera margaritifera*), which is an endangered species (Moorkens *et al.* 2017). In their life cycle, freshwater mussels have a specialized larva (i.e. glochidium), which must usually parasitize a host fish upon which it encysts and metamorphoses. *Salmo trutta* and *S. salar* are the specific host fish on which the *M. margaritifera* glochidia develop. This species is an integral component of the river ecosystem, filtering particles, serving as a substratum or habitat for algae and benthic invertebrates, respectively. Therefore, these ecological functions performed by freshwater mussels can be compromised in regions with lower mussel specific richness as the Iberian Peninsula (see Fig. 7 in Lopes-Lima *et al.* 2017), highlighting the conservation interest of its host species.

Brown trout, outside its native distribution range is involved in different drawbacks with native fish species such as (1) reducing population size by different ways (e.g. food competition, predation), (2) hybridization (e.g. with *Salvelinus fontinalis* in United States of America, USA; see Blanc and Chevassus 1986; Sorensen *et al.* 1995). Being this species considered by the Invasive Species Specialist Group (ISSG) as one of the 100 World's Worst Invasive Alien Species (http://www.issg.org/worst100_species.html).

1.1.5 Threats to brown trout

Brown trout is as Least Concern in the IUCN Red List category (Freyhof 2011); however, it is affected by habitat fragmentation, water pollution, global warming, overfishing and hatchery introgression.

One characteristic in any ecosystem is the connectivity. Referred to river ecosystems there are many definitions (see Table 1 in Wohl 2017). One of them would be the circulation of organisms, matter, and

energy through the riverine landscape *via* the aqueous medium (Amoros and Roux 1988). As a multidimensional space, rivers have different connectivity components (i.e. vertical, lateral, and longitudinal), nevertheless for salmonids the most studied was longitudinal connectivity. The most prominent constructions in rivers are dams, built for different human purposes (e.g. hydroelectric power, water level regulation and irrigation). Nevertheless, dams affect fish migration, isolating populations, and can affect water properties and habitat conditions, upstream and downstream (e.g. mean summer temperature in downstream water; see Lessard and Hayes 2003). Heggenes and Røed (2006) detected a significant correlation between years since dam construction and F_{ST} values, with an increasing local differentiation presumably caused by genetic drift. Birnie-Gauvin *et al.* (2018) documented the high increase of smolt number in Villestrup River (Denmark) after the removal of six dams with fishways. To improve river connectivity, nature-like bypass channels have been built in different rivers with good results, although they require an important investment (Dodd *et al.* 2017). For study and compare river connectivity, the Catalonian Water Agency developed in 2006 the *Index de Connectivitat Fluvial* (ICF; Solà *et al.* 2011).

Water pollution is the introduction of contaminants, related with human-mediated activities, in water bodies. Brown trout is considered a sensitive species to water pollution (see Luckenbach *et al.* 2001; Pickering 1989), although there have been reported some cases of tolerance in highly polluted ecosystems (e.g. metal pollution; Durrant *et al.* 2011), related to molecular mechanism of tolerance (see Webster *et al.* 2013). There is a wide variety of contaminants studied with the brown trout, from pesticides to microplastics (see Nusbaumer *et al.* 2020; O'Connor *et al.* 2020) and the bioaccumulation in different tissues due to the importance for human consumption (see Linde *et al.* 2004).

The ecological impacts of human societies have increased over the centuries, especially since the industrial revolution. These impacts are carried out in all scales (i.e. from local to global scale). A new geological epoch-term has been coined to reflect the magnitude of these impacts: “The Anthropocene” (Crutzen and Stoermer 2000). Global warming is one of the phenomena best known by society, with a range of consequences for the climate, economy, biodiversity, etc. The increase of temperatures would lead to the loss of suitable thermal brown trout habitats in lower latitudes (Almodóvar *et al.* 2012) or mountain regions (Hari *et al.* 2006). Although, adaption can play a role to changing temperature regimes (Jensen *et al.* 2008).

Restocking can endanger native populations by disrupting local adaptations with the introduction of domestication selection (Pinter *et al.* 2019) and influencing over different traits as dispersal (Saint-Pé *et al.* 2018). The current trends are conservation aquaculture approaches (e.g. enhance wild populations with individuals of the highest genetic similarity; see Froehlich *et al.* 2017) or the release of triploid individuals for fishing (i.e. sterile individuals). Nevertheless, triploidy is not always 100% successful and effective methods to detect diploid individuals would be necessary to avoid the release of thousands of fertile hatchery individuals (Sanz *et al.* 2020).

1.1.6 Human perspective

Recreational fishing is the main use of wild fish stocks in freshwaters belonging to developed countries (Arlinghaus *et al.* 2002), although it also exists in developing countries. According to the FAO, recreational fishing would be “fishing of aquatic animals (mainly fish) that do not constitute the individual’s primary resource to meet basic nutritional needs and are not generally sold or otherwise traded on export, domestic or black markets" (FAO 2012). When this activity implies a displacement outside the region of origin of anglers, it could be spoken about angling tourism, linked with active tourism,

ecotourism, and village tourism (FAO 2012). There are some studies about the economic impact of recreational fishing in some regions (e.g. Scotland, Butler *et al.* 2009; Gotland, Blicharska and Rönnbäck 2018), that can reach estimated incomes of the order of hundreds millions euros (see Inland Fisheries Ireland 2013). The number of freshwater fishing licences issued in Spain in 2018 was around 490,000 with an economic value close to seven million euros (Ministerio para la Transición Ecológica y el Reto Demográfico, MITERD 2020). To face up the angler's demand, every year releases of individuals or eggs of different freshwater species are performed. Among them, the brown trout represents the highest number of individuals or eggs released, exceeding 60% of the total (MITERD 2020). These data are not strange if we bear in mind that brown trout is one of the most valued species by anglers (FAO 2012), being one of the fish species providing an ecosystem service of utilitarian value (Millennium Ecosystem Assessment 2005).

1.2 BROWN TROUT GENOMICS

1.2.1 Whole genome duplication and diploidization of salmonids

Brown trout is a species of tetraploid origin in process of diploidization (Leitwein *et al.* 2017). Polyploidy can represent advantages or disadvantages for organisms. For instance, in sexual organism advantages would be related to heterosis and gene redundancy. The former could render more vigorous fish than their diploid counterparts, while the last would shield polyploids from the deleterious effect of mutations. On the other hand, polyploidy can originate problems in meiosis due to homologous mispairing and bad segregation, thus giving rise to gamete unbalance and lower viability (Comai 2005).

There have been four whole genome duplication (WGD) events from the common ancestor of vertebrates to salmonid divergence respect to the rest of teleost fish. Two whole genome duplications (referred to as the 1R and 2R WGDs) occurred in the common ancestor of all vertebrates, determining an increase in chromosome number from 10-13 in the protovertebrate karyotype to 40-52 in most vertebrates (Nakatani *et al.* 2011; Ohno 1970). A third (3R) WGD occurred 400 million years ago (Ma) in the ancestor of teleost fish, while a fourth salmonid-specific WGD (Ss4R) took place ~90 Ma in the ancestor of salmonid fish (Allendorf and Thorgaard 1984; Macqueen and Johnston 2014).

According to Glover *et al.* (2016), chromosomes or genes originated by whole-genome duplication events within species are called ohnologs, whilst those originated by allopolyploidization of a hybrid between different species would be homoeologs. Nowadays, the Ss4R is considered an autotetraploid genome duplication event (Campbell *et al.* 2020; Christensen and Davidson 2017). Salmonids would be in a diploidization process after the duplication event (Pouzadoux *et al.* 2017), which would imply disomic inheritance. This diploidization would not have been completed since multivalent pairing is observed at salmonid meiosis (Allendorf and Thorgaard 1984) supporting the tetrasomic inheritance observed (Allendorf *et al.* 2015), specially toward the telomeres (Nugent *et al.* 2017). The Atlantic salmon is the closest relative of brown trout and the number of chromosomes of their karyotypes largely differ (*S. salar* 2n = 58 *vs S. trutta* 2n = 80). There have been reported signals of intense chromosomal rearrangements through the evolution of these species, consistent in 13 fusions and two fissions which occurred in the Atlantic salmon branch after *S.salar*/*S.trutta* divergence (Pouzadoux *et al.* 2017).

Salmonids are gonochoristic (separated sexes), and sex would be genetically determined following a XX/XY system (Davidson *et al.* 2009; Yano *et al.* 2013). Nevertheless, sex determination can be influenced by environmental factors (e.g. water temperature), including the presence of pollutants or hormone treatments (Johnstone *et al.* 1978). X and Y chromosomes in salmonids display a large pseudoautosomal region and a small sex determining region (Davidson *et al.* 2009), so chromosomes are essentially homomorphic when inspected with optical microscope. In the last years, a novel master sex-determining gene has been characterized in rainbow trout (Yano *et al.* 2012), named sdY (sexually dimorphic on the Y chromosome). The presence of sdY only in phenotypic males of most salmonid species strongly suggests its conservation in the group (Araneda *et al.* 2019; Yano *et al.* 2013).

1.2.2 Genetic structure and phylogeography

Brown trout is a species with high genetic differentiation among populations, even those belonging to nearby sections of the same river (Bouza *et al.* 1999; Fernández-Cebrián *et al.* 2014), being one of the most structured vertebrate species (F_{ST} > 0.60; Ferguson 1989). However, this general rule is not always fulfilled (Heggenes and Røed 2006). In the Iberian Peninsula, this high differentiation was observed between river basins of the Mediterranean and Atlantic slopes using different molecular markers, such as allozymes (Bouza *et al.* 1999, 2001), Restriction Fragment Length Polymorphisms (RFLPs; Machordom *et al.* 2000) and microsatellites (Campos *et al.* 2006; Fernández-Cebrián *et al.* 2014; Martínez *et al.* 2007; Vera *et al.* 2010).

Brown trout displays a complex phylogeography, tangled by human-induced processes of introgression between lineages (Sanz 2018). Five major mitochondrial lineages were initially identified in brown trout using the control region of mitochondrial DNA (CR mtDNA; Bernatchez *et al.* 1992): Adriatic (AD), Atlantic (AT),

Danubian (DA), Marmoratus (MA) and Mediterranean (ME). Since then, several studies reported additional lineages restricted to the Iberian Peninsula, the Duero lineage (DU; Vera *et al.* 2010); the Tigris lineage in Turkey (TI; Sušnik *et al.* 2005); and an eighth North African lineage (Tougard *et al.* 2018), previously considered a different species (i.e. *Salmo macrostigma*). The mtDNA control region has been the main molecular marker used to identify geographically structured phylogenetic assemblages (Fig. 5) and mtDNA haplogroups have been used as reliable proxies of evolutionary significant units (ESUs) combined with nuclear molecular markers such as restriction fragment length polymorphisms (RFLPs; Castro *et al.* 1999; Schöffmann *et al.* 2007), internal transcribed spacers (ITS) of rDNA (Presa *et al.* 2002), microsatellites (Vera *et al.* 2013) or single nucleotide polymorphisms (SNPs; Pustovrh *et al.* 2011).

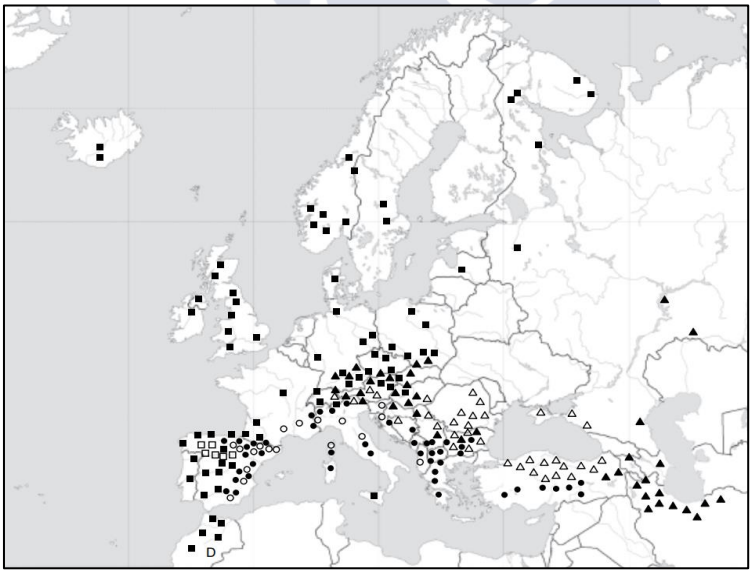

Figure 5. Geographical distribution of the *S. trutta* species complex lineages based on mitochondrial DNA data. Four lineages are represented in the Iberian Peninsula: Atlantic (black squares), Duero (Iberian Peninsula endemism; white square), Adriatic (black circle) and Mediterranean (white circle). Figure from Sanz (2018).

1.2.3 Hybridization and introgression scenarios

Within fish, hybridization is facilitated by reproductive features such as external fertilization and common spawning grounds (Hubbs 1955). When talking about hybridization and introgression in brown trout three different scenarios should be highlighted: (1) genetic introgression of wild populations with hatchery stocks, (2) natural hybridization between evolutionary lineages (secondary contacts), and (3) hybridization with closely related species (e.g. *S. salar*).

Releasing farmed brown trout into the wild has been a management practice applied since the early 1900's. Stocking in European river basins have been mainly performed with a stock of Central European origin (Martínez *et al.* 1993; Morán *et al.* 1991; Vera *et al.* 2018). This practice has been performed with other salmonid species as well (e.g. *Salvelinus fontinalis*, Lehnert *et al.* 2020; *S. salar*, Saltveit 2006).

The second aforementioned scenario is related to natural hybridization between lineages and/or subspecies, even though the validity of the *Salmo trutta* complex taxonomy is controversial. Experiments performed to test reproductive isolation between *S.trutta* subspecies belonging to the same mitochondrial lineage finally demonstrated a single biological entity (four subspecies belonging to Danubian lineage; Kalayci *et al.* 2018). In the Iberian Peninsula two main natural hybridization scenarios have been reported: (1) between AT and DU lineages in Duero River Basin and Miño-Sil River Basin with a parapatric distribution pattern (Martínez *et al.* 2007; Vera *et al.* 2010; Vilas *et al.* 2010); and (2) between AD and MED with a mosaic distribution pattern (Sanz *et al.* 2002).

The final scenario outlined before would include: (1) natural interspecific hybridization, and (2) human-mediated interspecific hybridization. In the first one, probably the most relevant case would be the hybridization between *S. trutta* and *S. salar* in the north of Spain

where both species live in sympatry. An average hybridization rate of 2.3% was reported by Leaniz and Vespoor (1989). The second one would be when two salmonid species did not originally live-in sympatry, being one of them non-native species. Different cases were reported between *S. trutta* and *S. farioides* (Škraba *et al.* 2018) and *S. fontinalis (*Cucherousset *et al.* 2008*).*

There are two hierarchical levels when estimating ancestry components in admixed populations under gene flow between different genetic clusters: (1) global ancestry and (2) local ancestry. Global ancestry inference estimates the overall proportion from each contributing population, over the individual admixed genome (Alexander *et al.* 2009), with a resolution and consistency depending on the molecular markers used for the analysis. This approach is sufficient to estimate ancestry in individuals and populations that are subjected to these gene flow processes. There are different software for these analyses following different statistical models. One of the most used is STRUCTURE (Pritchard *et al.* 2000; with more than 30,000 citations), based in linkage disequilibrium (LD), the non-random association of alleles, among unlinked loci and departures from Hardy-Weinberg proportions to define population structure. Nevertheless, new methods have been developed over the last two decades (see Table 1 in Porras-Hurtado *et al.* 2014). Local ancestry inference estimates the number of allelic copies from a particular source population at a particular genomic location (Thornton and Bermejo 2014). There are different software and models to deal with this approach, which can be classified in two main groups: (1) LD-based models and (2) non-LD-based models (Geza *et al.* 2019). Models belonging to the first group can be divided among those which account for: different types of LD (e.g. MULTIMIX) or uniquely admixture LD (e.g. STRUCTURE). Admixture LD would be caused when ancestry in nearby markers is inherited together due to admixture process. Non-LD-based local

ancestry inference models use other strategies as Bayesian (e.g. LAMP), conditional random fields (CRFs; RFMIX) approaches. Local ancestry software can work with phased or unphased data with different biological and statistical parameter requirements (see Table 1 in Geza *et al.* 2019).

1.3 CONSERVATION GENETICS/GENOMICS

Biodiversity levels

According to the IUCN biodiversity includes, on a decreasing scale, three fundamental levels: (1) ecosystem diversity, (2) species diversity and (3) genetic diversity. Despite, these levels have their own measurements, proposals for measuring biodiversity across them have been suggested, although genetics represents the keystone of these three levels (see Gaggiotti *et al.* 2018). Genetic variability represents the raw material for evolution and adaptation to environmental changes. Genetic diversity of each species can influence into community dynamics, specific structure (Lankau and Strauss 2007) and fluxes of nutrients and energy, resistance to disturbances at the ecosystem level (Hughes and Stachowicz 2004; Hughes *et al.* 2008). Roughly, conservation genetics has to do with the application of genetics principles to assist the conservation of populations, lineages, or species. This field encompass a series of actions such as: genetic management of populations to maximize retention of genetic diversity, definition of management units (MUs), resolution of "taxonomic knots", detection of hybrid individuals and zones between biological entities, forensic genetics to investigate wildlife crimes and traceability of products on the market. Nowadays, the importance of the genetic factors in the viability of species, lineages or populations is out of scientific debate.

Small populations

A decrease in effective population size, below a certain threshold, determines progressive inbreeding and a change in the genetic drift *vs* selection balance in favour of the former factor (Ellstrand and Elam 1993). It should be highlighted those genetic processes in small populations depend on the effective population size rather than on the real number of individuals of a population (i.e. census size). The effective population size (N_e) is the size of an ideal population with equal proportion of males and females where all contribute equally to the next generation, losing genetic diversity and increasing inbreeding at the same rate as the actual population (Frankham *et al.* 2014). The effective size of a populations is usually lower than the census size. In the case of populations with small effective population size both the increase of inbreeding and the loss of genetic diversity are concomitant.

Inbreeding is the production of offspring from individuals with common ancestors, which determines a decrease in heterozygosity due to the occurrence of homozygotes identical by descent, which often determines a fitness reduction due to the so-called inbreeding depression (lower survival and reproductive performances of offspring; Hays and Fagan 2016). Inbreeding is an inevitable consequence of small population size because in such scenario individuals have some kinship degree. Species exhibiting inbreeding depression in the wild have been reported, among others, within salmonids (see Wang *et al.* 2002). A factor to consider when studying inbreeding should be the chromosomic constitution of the species under study. The rate of increase in homozygosity is slower in polyploids than in diploids, thus polyploids are expected to suffer somewhat less inbreeding depression. A critical stage is when a population turns too small for different reasons (impacts related to human activities, demographic, or environmental stochasticity) becoming more inbred and

demographically unstable entering into a feedback cycle that is referred to as the extinction vortex (Fig. 6).

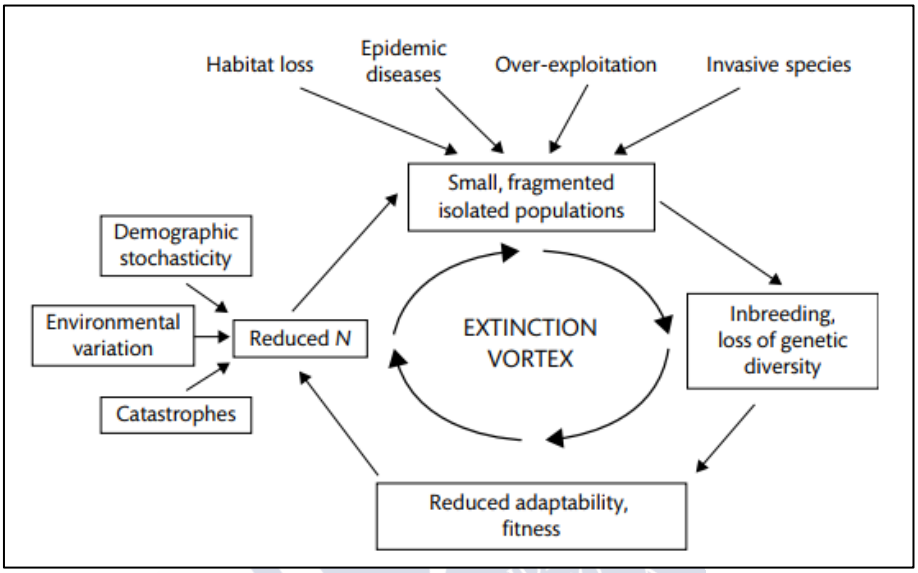

Figure 6. Extinction vortex with factors, feedbacks, and consequences. Figure from Frankham *et al*. (2010).

In large populations, mutation, selection, and migration have essentially deterministic effects, and chance have generally a minimal impact. Conversely, in small populations, the role of chance predominates, and the effects of selection are reduced due to genetic drift. Chance introduces a stochastic element into the evolution of small populations. Genetic drift is the process when alleles frequencies drift, from one generation to the next, within populations due to random sampling of gametes. Genetic drift tends to reduce especially allelic richness, but also heterozygosity (i.e. genetic diversity) within populations (Luque *et al.* 2016). Similar to inbreeding, the loss of genetic diversity (H_e) is higher in diploid than in tetraploid genomes (Wright 1969). Random sampling of gametes within small populations has three main consequences in evolution and conservation: (1) loss of

genetic diversity and fixation of alleles within populations with subsequent reduction in their ability to adapt and evolve, (2) divergence of allele frequencies for the same loci among populations from the same original source (e.g. fragmented populations), and (3) lower effect of natural selection.

Genetic diversity in populations is generated by mutations or introduced by migration. They are the only mechanisms for restoring the loss of genetic diversity; since mutation rates are usually very low for coding regions, this factor is negligible for recovering genetic diversity in endangered populations or species. Mutation is the ultimate source of all genetic variation. Threatened populations usually have lower genetic diversity than non-threatened populations (Frankham *et al.* 2010).

From conservation genetics to conservation genomics

According to Shafer *et al.* (2015), conservation genetics uses genetic markers (e.g. allozymes, microsatellites) to help conserving biodiversity and managing species and populations, whereas conservation genomics would use genome-wide information from high-throughput sequencing technologies (see next section) to achieve the aforementioned objectives. In the last fifteen years, due to sequencing advances, the number of molecular markers increased by several orders of magnitude (from dozens to thousands or millions), allowing systematic screening of populations on a whole-genome scale (Zhang *et al.* 2011). Nowadays, a broader range of population genetics questions can be addressed with genomics than before (Fig. 7), thus increasing the precision of classic analyses (e.g. population structure definition; Waples *et al.* 2020). The number of reference genomes available has sharply increased and new initiatives are emerging to obtain thousands of new reference genomes as Ten Thousand Fish Genome Project (Fish10K; Fan *et al.* 2020), to sequence 10,000

representative fish genomes within 10 years and more outputs from different genome consortiums (see Galla *et al.* 2019).

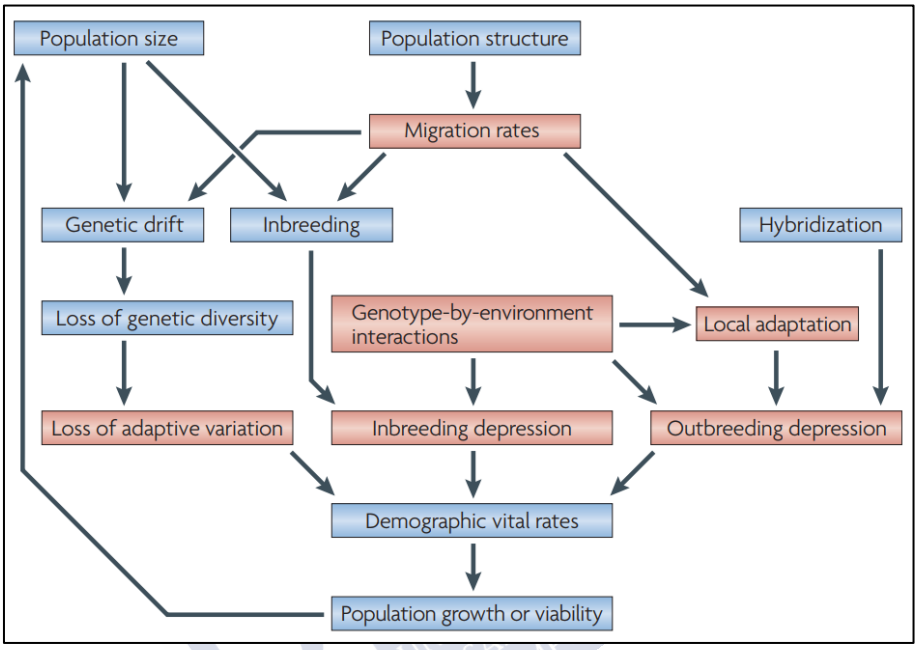

Figure 7. New horizons with conservation genomics. Conservation genetics with neutral markers (e.g. microsatellites) provides direct estimates of some questions (blue). Conservation genomics can address a wider range of questions (red). Diagram from Allendorf *et al.* (2010).

1.4 GENETIC AND GENOMIC TOOLS

1.4.1 DNA sequencing technologies

DNA sequencing is the process of deciphering the nucleotide composition on DNA strands. To date, four new sequencing generations have occurred (Mignardi and Nilsson 2014). First generation sequencing (FGS), where Sanger method (i.e. dideoxy sequencing) was the dominant technique (Wong *et al.* 2019), for thirty years (Schuster 2008). Sanger protocol is based on dideoxi-nucleotides which lack an hydroxil group in C3' thus determining the stop of the polymerization.

Dideoxi-nucleotides are labelled with different fluorescent labels, activated by the laser into automated sequencers (Freeland 2011). Second-generation sequencing (better known as Next Generation Sequencing, NGS) is a high-throughput DNA sequencing protocol where billions of DNA strands are sequenced in parallel after an amplification step of each individual DNA molecule. This technology has reduced prices drastically and generated an enormous amount of data (reads: sequences of nucleotides from the sequenced DNA molecules). This volume of information requires new bioinformatic tools and more hardware requirements (e.g. multicore processors, higher volume of memory), that can be achieved easily in supercomputing centres (e.g. BSC, Barcelona Supercomputing Center; CESGA, *Centro de Supercomputación de Galicia*). Third generation sequencing (TGS) does not require amplification and single DNA molecules are sequenced in real time (SMRT; Wong *et al.* 2019). Recently, fourth-generation sequencing has been described as the compendium of techniques to conduct genomic analysis directly in the cell and this is expected to be useful in more specific applications than other sequencing generations with broader applicability (Ari and Arikan 2016).

Since the beginning of NGS technology, different platforms have emerged with specific sequencing features such as read length, number of DNA molecules sequenced in parallel, etc (von Bubnoff 2008). Related to read length, there are two broad categories: Short-read NGS (usually < 300 pb) and long-read NGS (> 10 kb on average) (Mantere *et al.* 2019). Short-read NGS can be classified in two general types: sequencing by synthesis (SBS) and sequencing by ligation (SBL) (Slatko *et al.* 2018). Briefly, SBS uses DNA polymerase to incorporate complementary nucleotides to the elongating strand (Illumina platform) and SBL uses the mismatch sensitivity of DNA ligase to determine the underlying sequence of nucleotides in each DNA sequence (SOLiD

platform; Ho *et al.* 2011; Huang *et al.* 2012). According to Goodwin *et al.* (2016) the sequencing by synthesis can be classified as: cyclic reversible termination (CRT; Illumina, Qiagen platforms) and single-nucleotide addition (SNA; 454 pyrosequencing devices, Ion Torrent). In this thesis a sequencing by synthesis approach was used, provided by Illumina technology, which will be explained in more detail. For more information, it is encouraged the reading of these reviews: Goodwin *et al.* (2016) and Wong *et al.* (2019).

Illumina sequencing workflow can be summarized in three steps: (1) DNA library preparation by fragmentation of DNA followed by 5' and 3′ adapter ligation with barcodes to identify individuals that are sequenced together (multiplexed); (2) amplification of each fragment using bridge-PCR in a flow cell, which generates clonal clusters from bound fragments; and (3) read sequencing with fluorophore-labelled nucleotides, terminally blocked with a 3'-O-azydomethyl group, which hybridize to complementary bases linked to the emission of a fluorescent colour. Finally, fluorophores are washed, and nucleotides 3'-OH group are regenerated, by this way a new cycle begin with the addition of new nucleotides. This cycle is repeated "n" times to create a read length with "n" nucleotides. Sequencing data is exported to an output file, usually FASTQ format with the nucleotide sequences of each read with a measure of the quality of the nucleotide identification.

Once, the DNA of samples is extracted and prepared, different genomic approaches can be followed to prepare the DNA library. On one hand, the whole genome sequencing (WGS) procedure determines the nucleotide order in the target genome in one run (genome re-sequencing). Nevertheless, this option was until recently prohibitive on a population scale (Luca *et al.* 2011). On the other hand, using reduced representation libraries (RRLs) to work with a fraction of the whole genome in a more cost-effective way. With this approach reduced subsets of loci can be obtained with exome capture, transcriptome

sequencing (i.e. RNA-seq), etc (Hirsch *et al.* 2014). A subset of loci can be obtained through restriction enzyme digestion, followed by high-throughput sequencing of genomic sequences adjacent to the enzyme cut-site (restriction site-associated DNA sequencing, RAD-seq). A variety of RAD-seq techniques has been developed (2bRAD, ddRAD, etc; see Andrews *et al.* 2016) from the original publication (Baird *et al.* 2008). Some reasons that fuelled the popularity of RAD-seq were its feasibility for different genomes and the non-need of a reference genome, enabling studies of non-model organisms (Davey *et al.* 2011).

1.4.2 Molecular markers

Molecular population genetics was born 50 years ago with the first measures of genetic variation in allozyme loci (i.e. protein variants), the first molecular markers (Casillas and Barbadilla 2017). Molecular markers in sexually reproducing organisms can have biparental inheritance (nuclear DNA) or uniparental inheritance (mitochondrial DNA and plastids). Molecular markers are usually co-dominant meaning higher accuracy and informativeness, enabling to obtain genotypic and allelic frequencies in populations, as the basic information for their later application using the statistical methodologies of population genetics.

Microsatellite loci are tandem repetitive DNA of two to six base pairs motifs occurring at different genomic locations in a high frequency (on average 1 microsatellite/30 kb). SNPs could theoretically have a maximum of four possible variants, although nearly all SNPs are biallelic due to the low mutation rate per nucleotide site (Brown 2018). Microsatellites show much higher variation due to their much higher mutation rate (10^{-3} - 10^{-5}; Castro *et al.* 2006) thus having a much higher number of alleles per locus. Nevertheless, SNPs are much more frequent than microsatellites throughout the genome (1 SNP/300 bp in the human genome), thus accounting for approximately 90% of the total genetic variation (Collins *et al.* 1999).

To date, the markers used to estimate different population metrics in Spanish brown trout populations (allozymes, microsatellite loci or mitochondrial DNA control region; see Garcia-Marín *et al.* 2018) are assumed to be selectively neutral, reflecting only demographic processes from the balance between gene flow and genetic drift.

1.5 BROWN TROUT POPULATION GENOMICS

Different studies using genomic strategies have been reported to date in European brown trout, some of them using SNP-arrays (see Bekkevold *et al.* 2019; Linløkken *et al.* 2017). The diverse topics addressed are related to assessing genetic structure and identification of conservation units, restocking evaluation and monitoring, detection of adaptive variation (i.e. signatures of selection) in a wide range of features (e.g. individual growth) and linkage mapping, among others. Linløkken *et al.* (2017) using a 3,781 SNP panel detected outlier SNP loci under selection linked to genes involved in growth. Bekkevold *et al.* (2019) with a similar number of SNPs refined population structure and identified signatures of selection and their relationships to environmental factors. Leitwein *et al.* (2018) with 75,684 SNPs evaluated the impact of restocking from domestic Atlantic and Mediterranean lineages into supplemented wild populations across the genome using a local ancestry approach. Lemopoulos *et al.* (2018) with ~20,000 SNPs identified eight candidate genes potentially associated with brown trout migratory behaviour. Leitwein *et al.* (2017) with about 20,000 SNPs constructed a linkage map for brown trout to study chromosomal rearrangements, recombination rates and the effect of selection on neutral diversity. Bernaś *et al.* (2020) with 3,843 SNPs studied population structure and detected outlier loci in Southern Baltic region. Finally, Jacobs *et al.* (2018) analysed the evolutionary history between two reproductively isolated life history morphs in Scotland obtaining genes potentially related to observed phenotypic differences.

To our knowledge, no research following a population genomics approach has been addressed to date with brown trout in the Iberian Peninsula.

Until 2015, the main way to build loci and call genotypes with bioinformatic tools in brown trout using RAD-seq was a *de novo* approach (i.e. without reference genome). The genome of Atlantic salmon (*Salmo salar*), a congeneric species, was available since 2015 from the International Cooperation Project to Sequence the Atlantic Salmon Genome initiative (ICSASG). Reference genome approaches were carried out in brown trout RAD-seq data using the Atlantic salmon as reference genome (see Lemopoulos *et al.* 2018, 2019b; Paris *et al.* 2017; Valette *et al.* 2020). While reference genomes of closely related species can be used to obtain SNPs to estimate different population genetics parameters (e.g. nucleotide diversity; see Galla *et al.* 2019), accurate requires the species' own reference genome. For instance, conspecific genomes are recommendable to obtain the highest possible resolution (i.e. number and position of SNPs) to study patterns of hybridization along genome. In June 2019 was available at National Center for Biotechnology Information (NCBI) the first brown trout genome. This achievement was made within the framework of the "25 genomes for 25 years" initiative and promoted by the Wellcome Sanger Institute (https://www.sanger.ac.uk/collaboration/25-genomes-25-years/; Hansen *et al.* 2021). Because it has been available for just over a year, there are very few published research using brown trout reference genome (see De-Kayne *et al.* 2020; Grimholt and Lukacs 2021; Sudhagar *et al.* 2020).

2 OBJECTIVES

The present thesis aimed at studying the genetic structure of brown trout (*Salmo trutta*) in the Atlantic and Mediterranean slopes of the Iberian Peninsula using a population genomics approach. Natural and artificial hybridization and introgression patterns were studied and all the information was considered for the management and conservation of native genetic resources of this species in the Iberian Peninsula. The specific objectives of the study were:

1. Evaluation of different bioinformatic pipelines to identify and genotype SNPs to achieve consistent biological interpretations through a population genomics approach.

2. Analysis of genomic introgression of wild populations by hatchery stocks in the unstable ecosystems of the Mediterranean drainage.

3. Preliminary study of natural hybridization in a secondary contact in the Atlantic slope.

4. Improvement of conservation and management strategies to maintain the integrity of native lineages and recover natural populations in areas with high introgression rate due to restocking.

3 LOW IMPACT OF DIFFERENT SNP PANELS FROM TWO BUILDING-LOCI PIPELINES ON RAD-SEQ POPULATION GENOMIC METRICS: CASE STUDY ON FIVE DIVERSE AQUATIC SPECIES

3.1 INTRODUCTION

Next-generation sequencing technologies have represented a breakthrough for genomic studies (Quail *et al.* 2012) due to the huge reduction of sequencing cost (less than 0.02$ per Mb; Wetterstrand 2020) and the development of a broad and versatile range of techniques for different genomic approaches (Goodwin *et al.* 2016). By harnessing the possibilities of NGS, diverse reduced-representation genome sequencing approaches, useful to identify and genotype thousands of markers for genomic screening, were suggested and quickly became popular (Baird *et al.* 2008; Davey *et al.* 2011). One of these approaches is the restriction site-associated DNA sequencing (RAD-seq), currently in a more mature phase, which includes different methods (e.g. ddRAD-seq, ezRAD-seq, 2bRAD-seq) whose performances have been compared using simulations and real data (Andrews *et al.* 2016). RAD-seq methods require specific library preparation protocols, which exploit the ability of restriction enzymes (REs) to cut at specific genomic targets rendering a collection of fragments representative of a

genome fraction to be compared among samples. These collections can be screened to identify and genotype a variable number of SNPs depending on the goals of the study for population genomics, linkage mapping or genome wide association studies, among others. The 2b-RAD method here used exploits the properties of IIB REs which produce a collection of short DNA fragments (between 33 and 36 bp) by cutting at both sides of the recognition site (Wang *et al.* 2012). This method has the advantages of simple library preparation, short-reads to be sequenced (single-end 50 bp) and, as other methods, the number of loci can be adjusted both using REs with different recognition site frequency or by fixing nucleotides in the adaptors during library construction (i.e. selective-base ligation) (Barbanti *et al.* 2020; Wang *et al.* 2012).

Genomic laboratory protocols have been set up and optimized through years by introducing modifications on the original RAD-seq methodology to get better results using different laboratory protocols for different scenarios (e.g. samples with low DNA quality, genome size, etc; see Fig. 5 in Barbanti *et al.* 2020). Similarly, the bioinformatic pipelines starting from raw data, a critical issue in RAD-seq methodologies, have undergone an important refinement and diversification. Nevertheless, there is not a consensus about what is the best strategy for each scenario, despite the increasing number of studies addressed to evaluate the impact of technical and/or bioinformatic protocols (Díaz-Arce *et al.* 2019; O'Leary *et al.* 2018). In a typical 2b-RAD library, hundreds of millions of reads are generated, and they need to be allocated to each multiplexed individual (dozens to hundreds in the same lane) and to each genomic position or locus in the reference genome (or RAD-tag catalogue). The rationale behind this is stacking raw reads belonging to the same locus, while discerning and separating at the same time the reads belonging to different loci. Results could be improved if a reference genome, belonging to the species itself or to

other congeneric species, is available. This would enable to avoid mixing of reads pertaining to paralogous loci. In November 2020, there were reference genomes for 25 bivalve species and subspecies (22 genera) and 583 fish species (338 genera) with different assembly confidence at the NCBI database (https://www.ncbi.nlm.nih.gov/datasets/). Nevertheless, there are about 9,200 species within the 1,260 bivalve genera (Huber 2010) and 35,672 recognized species within the 5,212 documented fish genera (Fricke *et al.* 2020). All in all, less than 0.2% of the genomes of the known eukaryotic species have been sequenced to date (Lewin *et al.* 2018). Although full genome sequencing assembly is becoming progressively more robust thanks to the long-read sequencing methods and assembling strategies, most of the species will have to wait for long before their genomes are assembled. Therefore, *de novo* approaches (i.e. stacking reads without a reference genome) will be the only option for many studies, although some initiatives are trying to change this perspective (e.g. Earth Biogenome Project; https://www.earthbiogenome.org/). For this reason, one of the strengths of a RAD-based method is its applicability without a reference genome (Rochette and Catchen 2017).

There are different bioinformatic pipelines to identify a high number of SNPs and achieve confident genotypes using a RAD-seq approach. The most popular one is Stacks (Catchen *et al.* 2011, 2013) with around 3,000 citations, at Google scholar in Nov. 2020, but several other alternatives, including Meyer's 2b-RAD pipeline (Wang *et al.* 2012), which was the original building-loci pipeline for 2b-RAD data, have been recently published. Some of these pipelines are able to perform a *de novo* approach (dDocent; Puritz *et al.* 2014), whereas others need a reference genome for alignment (Fast-GBS, Torkamaneh *et al.* 2017; TASSEL-GBS v2, Glaubitz *et al.* 2014) or can address both approaches (Stacks, Meyer's 2b-RAD v2.1 pipeline, ipyrad; Eaton and Overcast 2020). Several of these alternative pipelines merge and

concatenate pre-existing applications, making their design flexible and customized according to the data managed and the goals of the study, but also providing upgrading and reliable bug-fix (e.g. dDocent, Fast-GBS, Meyer's 2b-RAD v2.1 pipeline). Several factors should be considered for the selection of the bioinformatic pipeline to be used, among which sampling variance (number of samples and reads across them), population structure, genome architecture of the species studied and budget (e.g. read coverage) are the most relevant. The genome of each species has its particular size, history (e.g. duplication events), polymorphism, complexity and interindividual variability, which can hinder the identification of stacks of reads (putative RAD-loci) and their variants, circumstances that should be considered when choosing the appropriate building-loci pipeline and its parameters.

Studies comparing bioinformatic pipelines and strategies already exist. Some comparisons between *de novo* and reference-based approaches are available (Shafer *et al.* 2017; Torkamaneh *et al.* 2016), and one of them tested the performance of the different strategies used to obtain accurate population genetics inferences (Shafer *et al.* 2017). Noticeable differences were observed among bioinformatic pipelines in the number of detected SNPs, sometimes resulting in distinct values for population descriptors and inference (Shafer *et al.* 2017). Other studies have evaluated the same software with different species to optimise the selection of bioinformatic parameters (Stacks 1.42; Paris *et al.* 2017); (Stacks 1.44; Díaz-Arce *et al.* 2019), making a common advice of doing preliminary trials to optimize the building-loci pipeline selected parameters. Published step-by-step protocols with a single species also exist (Rochette and Catchen 2017). A number of SNP calling comparison between Stacks 1.08 and dDocent 1.0 has been carried out using three fish species (Puritz *et al.* 2014), while Sovic *et al.* (2015) tested a novel pipeline (i.e. AftrRAD) *vs* Stacks and PYRAD using simulated and species datasets to assess computational efficiency and

SNP calling. It is not uncommon to find large differences in the number of SNPs (e.g. of one order of magnitude) in some building-loci pipelines comparisons (Puritz *et al.* 2014). Recently, Wright *et al.* (2019) compared population parameters (e.g. F_{ST}, PCoA) with SNPs obtained from three pipelines (i.e. GATK, SAMtools, Stacks) using two species with reference genome. Results showed remarkable differences in some population parameters (e.g. Hardy Weinberg Equilibrium) across bioinformatic approaches. Considering this information, a main issue that should be clarified on RAD-seq methodologies is the impact of building-loci pipelines on population genetics parameter estimations and derived conclusions using a *de novo* approach on different biological scenarios and to some extent, to be compared to a reference genome approach.

In this work, two building-loci pipelines for SNP calling and genotyping: (i) Stacks 2.0 (http://catchenlab.life.illinois.edu/stacks/) and (ii) Meyer's 2b-RAD v2.1 pipeline were tested using a *de novo* approach on different genomics and population genetics scenarios by using five aquatic species: (1) Manila clam (*Ruditapes philippinarum*), (2) common edible cockle (*Cerastoderma edule*), (3) brown trout (*Salmo trutta*), (4) silver catfish (*Rhamdia quelen*) and (5) small-spotted catshark (*Scyliorhinus canicula*). A range of population parameters were compared in the five species applying similar parameter settings for each pipeline (description of pipelines in Material and Methods). The two marine bivalve species from the Order Veneroida show high polymorphism and low population structure (Martínez *et al.* 2015; Vera *et al.* 2016); the brown trout belongs to the order Salmoniformes, which suffered a specific genome duplication event (Leitwein *et al.* 2017), and shows one of the highest population structuring among vertebrates (Ferguson 1989); isolated populations from different ecosystems were analysed in the silver catfish, a freshwater species from the order Siluriformes living in fluvial and costal lagoon environments (Ríos *et*

al. 2020); finally, the small-spotted catshark (order Carcharhiniformes) is a benthic species which populations here used show low genetic differentiation (Manuzzi *et al.* 2018). Despite these species represent different population genetics and evolutionary scenarios, this does not mean that they necessarily comprehend all the models for the manifold scenarios used to check the performance of building-loci pipelines. To date, two of the species used in this study have a reference genome available: Manila clam (assembly size: 1.123 Gb, 19 chromosomes; Yan *et al.* 2019) and brown trout (2.370 Gb, 40 chromosomes; see *S.trutta* assembly at NCBI https://www.ncbi.nlm.nih.gov/assembly/GCF_901001165.1). There are different statistics to evaluate the quality of genome assemblies (see Table 2 in Wajid and Serpedin 2016) such as scaffold N50 and N90 (i.e. the length of the scaffold at which 50 and 90% of the assembly length is covered, respectively). The scaffold N50 is much higher in brown trout than in Manila clam assembly (52,209 Kb and 345 Kb, respectively), but with a very high accuracy and completeness (see Yan *et al.* 2019 Supplementary Material). Anyway, for short read data such as for 2bRAD-seq, the contiguity of the genome should not have a major influence in the results when conservative short-read aligner parameters and strong SNP filtering steps are applied, minimizing the admixture of reads from paralogous loci or losing of reads due to genome fragmentation.

Wang and Guo (2004) hypothesized that bivalves with 19 chromosomes could have a tetraploid origin, due to its ability to tolerate chromosomal aneuploidies. Furthermore, gene/gene family expansions would be a rather common process in this group, likely more frequent than in other molluscs (Takeuchi *et al.* 2016). Both genomic features could suppose a challenge regarding paralogous genes for stacking reads, a common problem when no reference genome is available. In addition, molluscs present the highest genetic polymorphism in animal kingdom (Curole and Hedgecock 2007), which could represent

genotyping drawbacks related to the presence of null alleles. All salmonids, including brown trout, have a tetraploid origin in process of diploidization since their origin around 90 Ma. This specific duplication should be added to the three Whole Genome Duplications (WGDs) events in the line of teleosts from the vertebrate ancestor (Macqueen and Johnston 2014; Pasquier *et al.* 2016), which represent a major issue regarding paralogy. This issue would not be so important in small-spotted catshark and silver catfish, with two and three older WGD events in their evolutive lines, respectively (Berthelot *et al.* 2014; Donoghue and Purnell 2005). The most recent, the teleost-specific 3rd WGD, dated around 300 Ma. We followed a *de novo* approach for comparison between pipelines in all species, but reference genomes in these two species were also taken as a useful reference to elucidate which building-loci pipeline provides better results with a *de novo* approach.

For the five species, population parameters, structure pattern and outlier loci detection were estimated as an essential outcome to evaluate the performance of four SNP panels after different filtering steps. From two building-loci pipelines, Stacks (STA panel onwards) and Meyer's 2b-RAD v2.1 pipeline (ALT panel onwards), and two criteria for SNP selection, common SNPs (i.e. shared between building-loci pipelines, COM panel onwards) and merged SNP (a combination of shared and exclusive SNPs from both building-loci pipelines, MER panel onwards) panels. When available, the results from reference genome approach were used to compare the results of population parameters evaluated with the *de novo* approach. In this case three additional SNP panels were obtained: the former with the reference genome approach using Stacks and the two remaining with the shared SNPs (i.e. Stacks *de novo* and 2b-RAD v2.1). Our main goal was to assess the influence of the genomic architecture and population structure on the biological conclusions obtained with the different bioinformatic pipelines, and

accordingly, to propose methodological recommendations for future studies using a *de novo* approach.

3.2 MATERIAL AND METHODS

3.2.1 Sampling

All datasets used in the present study are own resources obtained from previous research carried out by the authors. Four Manila clam (*R. philippinarum*) localities, three from the Adriatic Sea (Italy: Chioggia, N = 30; Porto Marghera, N = 30; and Po River mouth N = 25; Milan *et al.* 2019) and one from the Atlantic Ocean (Spain: Galicia, N = 25), were studied. Four common edible cockle (*C. edule*) localities from the European Atlantic area (Somme Bay, France, N = 30; Campelo, Spain, N = 30; Miño, Spain, N = 30; and Ría Formosa, Portugal, N = 30) were used from regions with extractive cockle activities (Maroso *et al.* 2019). Three localities of brown trout (*S. trutta*) from Duero River basin in the Iberian Peninsula (Águeda, N = 15; Omaña, N = 20; and Pisuerga, N = 17), two of them representing different mitochondrial pure native lineages (Atlantic and Duero, Bouza *et al.* 2001; Vera *et al.* 2010) and one from the hybrid zone (Omaña; Martínez *et al.* 2007), were evaluated. Two localities of silver catfish (*R. quelen*), a Neotropical freshwater species distributed from the Northeast of Los Andes to the centre of Argentina and living in fluvial and coastal lagoon environments (Perdices *et al.* 2002), were analysed. These samples came from Sauce Lagoon (N = 10) and Uruguay River Basin (N = 11) belonging to two divergent lineages (Ríos *et al.* 2020). Finally, two nearby localities without genetic differentiation of small-spotted catshark (*S. canicula*) from North Sea (N = 13) and Irish Sea (N = 15) were analysed (Manuzzi *et al*. 2018). All information from samples analysed is summarized in supplementary material (Table S1).

3.2.2 Library preparation

DNA extraction and 2b-RAD libraries preparation using AlfI IIB RE followed the same protocol except for small-spotted catshark (Manuzzi *et al*. 2018) where CspCI IIB RE was used instead. The libraries were sequenced using Illumina sequencing platforms (i.e. HiSeq 1500 for small-spotted catshark, HiSeq 2500 for Manila clam and NextSeq500 for the remaining species) following a 50 bp single-end chemistry. For details see Manuzzi *et al.* (2018) and Ríos *et al.* (2020).

3.2.3 Bioinformatic analysis

Building-loci pipelines: background

Stacks 2.0 and Meyer's 2b-RAD v2.1 were the building-loci pipelines chosen for comparing their performances using a *de novo* approach within the broad genome and population genetics species scenarios selected. Meyer's 2b-RAD v2.1 pipeline and Stacks building-loci pipelines have some similarities on their strategy; roughly, both are based on stacking reads into putative loci by sequence similarity, assuming that each locus corresponds to a single place in the species genome. Nevertheless, there are many differences on how loci are built and how the user can control the existing options and genotyping strategies. Stacks, with a *de novo* approach, works firstly at the individual level demanding a number of identical reads to build a locus (i.e –m parameter in ustacks), while the 2b-RAD v2.1 pipeline works with a combined subset of samples to build a global reference panel to which align every read. For genotyping, Stacks uses a chi square test to call a heterozygote or a homozygote (i.e. –alpha and --gt-alpha), whereas nucleotide frequencies based on allele read depth at each position and sample are used for genotyping in the 2b-RAD v2.1 pipeline. Accordingly, huge differences in the raw SNP number were obtained in preliminary analysis in our study. Anyway, we tried to apply

the highest number of common parameters in both pipelines to be consistent among comparisons.

Stacks 2.0 pipeline can be summarized as follows (see Fig.1 in Rochette *et al.* 2019): (1) Raw sequence reads were demultiplexed and filtered according to different criteria such as quality, uncalled bases and read length (process_radtags); (2) reads from each individual were clustered into putative loci, and polymorphic nucleotide sites identified (ustacks); (3) putative loci were grouped across individuals and catalogues of RAD-loci, SNPs and alleles were constructed (cstacks); (4) putative loci from each individual were matched against the catalogue (sstacks); (5) the data were transposed to be oriented by locus, instead of by sample (tsv2bam); (6) all individuals were genotyped at each called SNP (gstacks); and (7) SNPs were finally subjected to population genetics filters (populations) and results written in different output files, e.g. Genepop; (Rousset 2008), STRUCTURE (Pritchard *et al.* 2000) file formats.

The *de novo* approach used for the 2b-RAD v2.1 pipeline can be summarized as follows: (1) from a subset of high quality reads (Phred quality scores ≥ 30 at all positions) from all samples, a global *de novo* reference catalogue was built by clustering these reads with the BuildRef.pl script and cd-hit 4.6.8 (Fu *et al*. 2012; Li *et al.* 2006), (2) every read was aligned against the reference catalogue using a mapping program (in our case Bowtie 1.1.2; Langmead *et al.* 2009); (3) allele frequencies were counted at each position (SAMBasecaller.pl) and genotypes determined from that information (NFGenotyper.pl); and (4) genotypes called across samples were combined into a single genotype matrix with samples as columns and loci as rows (CombineGenotypes.pl). Perl scripts belonging to the last version are publicly available (https://github.com/Eli-Meyer/2bRAD_utilities) and earlier versions on request.

A reference genome-based pipeline was used for Manila clam and brown trout, the species with chromosome assembly level reference genomes. In this case, the differences in the pipelines of Stacks 2 and Meyer's 2b-RAD v2.1 are lower. For instance, Stacks 2 reduces the number of modules necessary from six to two when using reference genome and the building of putative loci is conditioned by the step using a short-read aligner. Our goal was not to compare this option between pipelines, but to take as reference the genome-based approach to be compared with the *de novo* approach as a gold standard within each pipeline.

Building-loci pipelines: analysis

After demultiplexing raw data, several filtering criteria were applied: (1) all reads were trimmed and filtered by the RE recognition site to retain only those sequences of 36 bp (or 32 bp from CII RE catshark) centred on the RE recognition site using our own Perl scripts and Trimmomatic 0.38 (Bolger *et al*. 2014); and (2) process_radtags (module belonging to Stacks) was used to remove reads with uncalled bases (-c option). Other parameters were species-specific (e.g. window sliding size, -w; score limit, see Table S2A). To check raw and filtered reads quality, FastQC 0.11.7 (Andrews 2010) was used. Stacks input sequences were oriented in the same orientation using our own Perl script to avoid oversplitting. To say, the set of input reads for both building-loci pipelines in each species was always the same.

At the building-loci pipeline step, the parameters considered were: (1) minimum number of identical reads to create a stack (default values were used); (2) maximum number of mismatches between RAD-loci within and between individuals (-M 2/-n 2 for fish or -M 3/-n 3 for bivalves and their analogous parameters with ALT pipeline); (3) indels were discarded (i.e. --disable-gapped at different modules); (4) SNP calling model and alpha cut-off: their default values were used in the STA pipeline (Stacks 2.0), while in the ALT pipeline (Meyer's 2b-RAD

v2.1) we considered a range of 0.1–0.2 to determine the genotype at each position (default values are 0.01–0.25); to say, when the frequency of the less frequent allele was lower than 0.1 the genotype was called as homozygous while frequencies higher than 0.2 were called as heterozygous; intermediate allele frequencies for the less frequent allele were called as uncertain (see Tables S2B and S2C). When a reference genome was available, Bowtie 1.1.2 was used as short read aligner. The number of mismatches allowed between reads and the reference genome using the -v alignment mode was 2 mismatches for brown trout and 3 mismatches for Manila clam, the same as mentioned above. Only reads which aligned to a single site in the reference genome were considered (-m 1). The same parameter values were used in Stacks modules shared between reference-based genome and *de novo* approaches (i.e. gstacks and populations modules).

SNP filtering steps and creation of datasets

The raw SNP panels of the STA and ALT pipelines were filtered using the same parameters for consistency, retaining only a set of markers and alleles represented across the individuals genotyped. Filters were applied in the same order for each dataset (Fig. 8), as recommended (O'Leary *et al.* 2018): (1) Uniquely biallelic SNPs were included (BIAL filter); (2) a minimum locus coverage of eight reads was chosen (Min coverage filter); (3) RAD-loci with ≤ 3 SNPs were retained; (4) SNPs were retained only if the less frequent allele was represented at least three times at the whole species sample (MAC, minimum allele count, filter); (5) genotyped in at least in 60% of the individuals in each population for a SNP to be retained (POP filter); (6) SNPs were included when they adjust to Hardy-Weinberg expectations (p-value > 0.01) in more than half of the populations analysed (HW filter); and (7) only the first SNP per RAD-locus was retained when several SNPs were called in the same RAD-locus to avoid redundant information.

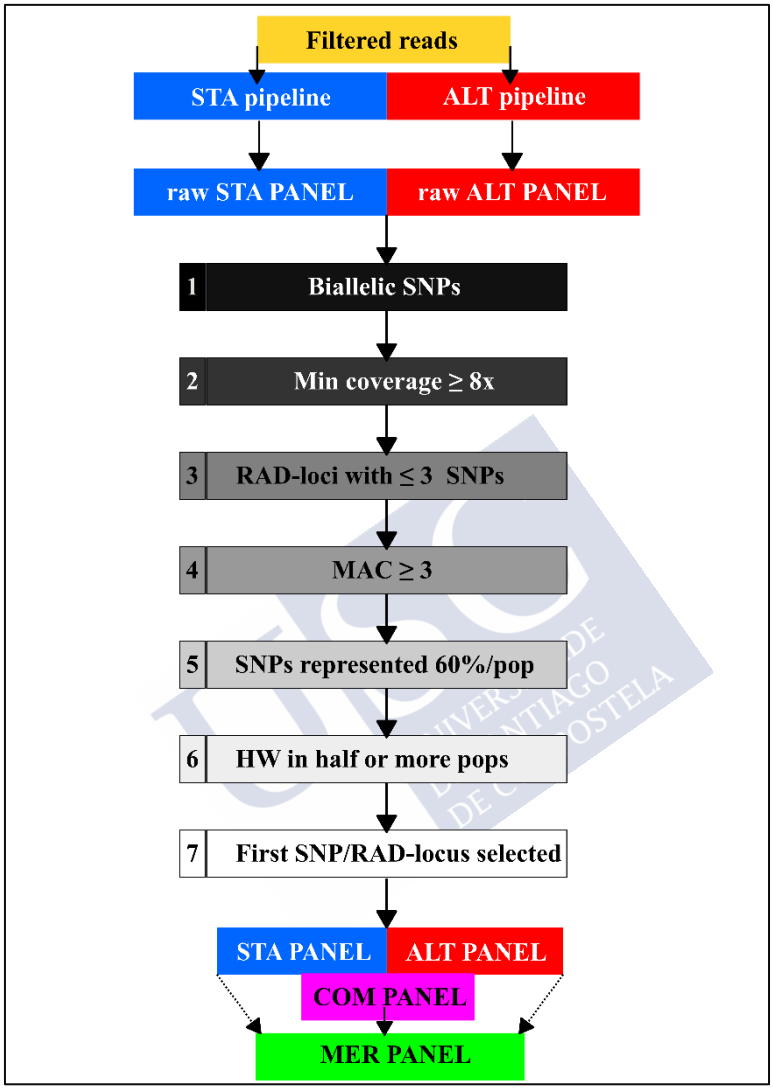

Figure 8. Scheme of filtering steps to obtain the different SNP panels. STA, ALT, COM and MER, representing Stacks, Alternative, Shared and Merged panels, respectively.

According to the aforementioned criteria, four SNP panels were tested and compared: (1) STA and (2) ALT SNP panels were further used to obtain (3) common (COM) and (4) merged (MER) panels.

When reference genome was available three additional SNP panels were obtained (i.e. RG, RG-STA, RG-ALT). RAD-loci of these panels from both pipelines were compared to identify shared and private RAD-loci. In order to do this, cd-hit-est was used to cluster similar RAD-loci taken from the two pipeline catalogues with the same threshold of similarity (-c) used in the clustering steps of building-loci pipelines (i.e. two mismatches maximum for fish species and three mismatches maximum for bivalve species). Furthermore, we used a -g value of 1 when clustering sequences to meet the established similarity threshold and a band alignment width (-b) of 1 to avoid previously separated sequences due to indels at specific clusters.

This procedure rendered COM and MER SNP panels created using a customized Perl script. SNPs at shared RAD-loci between STA and ALT panels were selected after the sixth filtering step, since the first SNP at shared RAD-loci could be different after the full filtering pipeline (Fig. 8). MER panel was finally created by taking the SNPs from "private" ALT and STA pipelines RAD-loci (i.e. those from cd-hit-est clusters with only STA or ALT pipelines RAD-tags) plus the COM SNP panel previously obtained. Again, only the first SNP per RAD-locus was retained to avoid redundant information. The Genepop files of all shared SNP panels were compared to quantify their genotyping differences using own Perl script (see https://github.com/abhortas/USC-RAD-seq-scripts).

Comparison of outputs and population genetics analyses

The results of the aforementioned pipelines were compared using both different quantitative (i.e. number of SNPs) and population genomics metrics. Filtered Genepop files were obtained using customized Perl scripts. These files were transformed for subsequent analyses using the PGDSpider 2.1.1.5 software (Lischer and Excoffier 2012). Firstly, the consequences of filtering over the number of RAD-loci and SNPs were evaluated for each combination of species-pipeline;

secondly, common, and private RADloci/SNPs between the two pipelines were obtained for each species. Finally, biological interpretations from each pipeline/species were compared, including basic population genetics results (i.e. genetic diversity levels and population structure).

Observed and expected heterozygosity (H_o and H_e, respectively), inbreeding coefficient (F_{IS}, using 1,000 bootstrap iterations to estimate their 95% confidence intervals) and allelic richness were calculated per population using diveRsity R package 1.9 (Keenan *et al.* 2013). Global F_{ST} calculation and HW tests were performed with Genepop R package 1.1.7 (Rousset 2008). STRUCTURE 2.3.4 software (Pritchard *et al.* 2000), using R package ParallelStructure 1.0 (Besnier and Glover 2013), was used to define the most likely number of population units (K) present with LOCPRIOR model with correlated allele frequencies model, testing K values from 1 to the number of sampling localities in the species dataset + 1 with 10 replicates composed by 100,000 Markov chain Monte Carlo (MCMC) replicates and a burn-in period of 10,000 steps. STRUCTURE results were parsed with STRUCTURE HARVESTER (Earl and vonHoldt 2012), which implements the Evanno's method (Evanno *et al.* 2005) to detect the most likely number of clusters according to the data. CLUMPAK (Kopelman *et al.* 2015) was used to merge runs with the same K that suggested similar patterns of structuring and to obtain cluster membership plots. As a second approach to detect population structure, a Discriminant Analysis of Principal Components (DAPC) analysis was performed based on genetic data, as implemented in R package Adegenet 2.1 (Jombart 2008; Jombart and Ahmed 2011). The optimal number of principal components to be used was estimated with the cross-validation method implemented in the package and from one to three discriminant components were retained according to the amount of population structure variation they explained. Finally, outlier loci potentially under

selection (OL), i.e. those showing higher or lower differentiation values (i.e. F_{ST}) across populations than the neutral background, were detected using the Bayesian approach implemented in BayeScan 2.1 (Foll and Gaggiotti 2008) with default parameters. Loci with a Log10 posterior odds (PO) higher than 1.5 were retained as potential outliers for later comparison among the four datasets resulting from the two pipelines.

3.3 RESULTS

The number of filtered reads loaded into building-loci pipelines using the reference genome approach was lower than with the *de novo* approach. A percentage of 22.1 and 25.4% were finally used in Manila clam and in brown trout, respectively. This reduction mostly due to those reads aligning to more than one place (59.6 and 72.6%, respectively) that were filtered out (i.e. -m 1 in Bowtie 1.1.2), the remaining reads failed to align with the mismatch criteria applied (18.3 and 2.0%, respectively).

The number of initial SNPs, after the building-loci step with the *de novo* approach, ranged from 56,074 in brown trout (STA) and 125,823 in silver catfish (ALT) to 356,389 (STA) and 426,317 (ALT) in common cockle (Tables S3-S7). These figures dropped throughout the successive filtering steps (Fig. 8) up to finally being retained from 0.2% in Manila clam STA panel to 20.5% in silver catfish STA panel (Fig. 9).

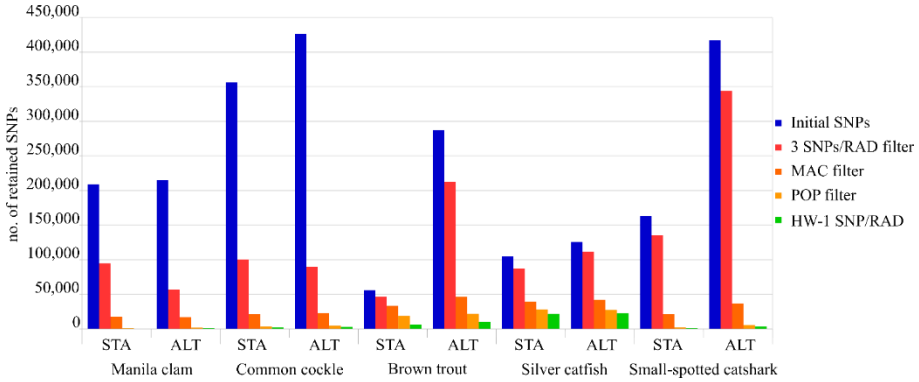

Figure 9. Number of SNPs from the initial building-loci pipelines (blue bars) to the final panels (green bars) through the different SNP filtering steps.

There was a remarkable difference at the initial number of SNPs obtained between STA and ALT pipelines in brown trout and small-spotted catshark, although the outcome after filtering was rather similar (Table 1). No comparison was made at BIAL filter (i.e. SNPs with more than two alleles excluded; Fig. 8), since triallelic SNPs are removed with Stacks by default. The proportion of missing genotypes after applying the minimum coverage filtering step was higher with ALT panel than with STA in almost all species (Fig. 10). Filtering patterns varied among species due to the different weight of each filtering step. For instance, in bivalves, where genetic polymorphism is higher, the SNP retention after the third filtering step (i.e. RAD-loci with ≤ 3 SNPs per RAD-locus) was much lower than in fish species (Fig. 9), while in silver catfish and small-spotted catshark, was due to the minimum allele count (MAC). This was related to the smaller sampling size of those fish species (N = 21 and N = 28, respectively) and the higher frequency of missing data, especially in small-spotted catshark (83% after depth filter in STA and ALT panels). When comparing pipelines, no clear differences on the filtering pattern were observed through the different filtering steps (Fig. 9), except for MAC in brown trout, where more SNPs were pruned in the ALT panel. This could be related to the

increment on the missing data after the minimum coverage filter, with higher frequency in ALT (59.3% *vs* 43.7% for ALT and STA, respectively (Fig. 10). After the third filtering step, in brown trout (N = 52) there were significantly more missing genotypes per SNP (*p*-value < 0.01) at ALT panel on average (28.74 ± 15.24) as compared to STA panel (21.07 ± 17.54). The two species with the lowest median coverage at this step were small-spotted catshark (median = 10x for ALT and STA panels), brown trout (median = 11x and 14x for ALT and STA panels, respectively) and Manila clam (21x for the STA panel). The final number of SNPs ranged from 479 (STA) and 956 (ALT) in Manila clam to 21,468 (STA) and 22,481 (ALT) in silver catfish. These figures were always higher with the ALT pipeline for all species. The number of SNPs in COM panels, those from STA panel shared with ALT panel, ranged from 206 in Manila clam to 17,459 SNPs in silver catfish, while the percentage of SNPs called in STA that were found in ALT ranged from 23.9% in small-spotted catshark to 81.3% in silver catfish. The main source of variation when considering the whole COM panels genotype dataset (for instance in silver catfish Ntotal COM genotypes (366,639) = Nsamples (21) x NCOM SNP (17,459)) was missing data, ranging from 2.6% in common cockle to 14.8% at small-spotted catshark (Figs. S1-S4). Roughly speaking, there would be three types of missing calling data with a *de novo* approach: (1) SNPs not genotyped by a building-loci pipeline due to too low coverage (< 3x with used configuration); (2) SNPs genotyped by a building-loci pipeline but with a coverage lower than 8x (Min coverage 8x filtering step; Fig. 8); and (3) SNPs with enough coverage but ambiguous alternative nucleotide depth (see http://eli-meyer.github.io/2bRAD_utilities/#genotype). The main source of missing genotype differences between pipelines was related to COM SNPs from STA pipeline that passed the minimum coverage filter (Min coverage 8x), but not were genotyped by ALT pipeline due to having a coverage lower than 3x. Both pipelines never genotyped with a

coverage lower than 3x with the used configuration. This situation was found in most species, causing from 46.8% of the total missing genotype differences between ALT and STA genotypes in shared SNPs (COM panel) in brown trout to 63.2% in small-spotted catshark.

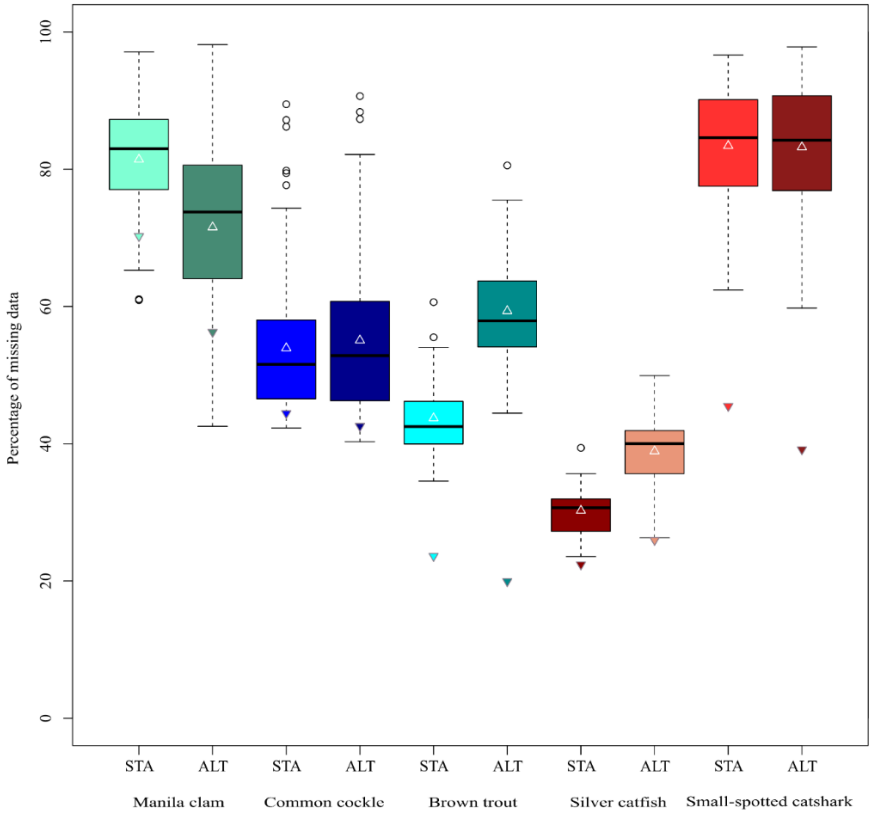

Figure 10. Percentage of missing genotypes after the Min coverage filter (8x). Boxplots were obtained with the percentage of missing genotypes through the different samples. Up and down triangles represent the percentage of missing genotypes at different SNP panels after and before 8x coverage filter, respectively.

Nevertheless, the main source of missing data differences in Manila clam was genotypes removed by coverage filter (Min coverage

8x) in ALT but not in STA homozygous-heterozygous differences between pipelines at the same SNP and individual ranged from 0.5% in brown trout to 2.6% in small-spotted catshark with respect to the whole COM genotype panel (Table 2). The frequency of genotyping differences caused by different homozygotes at the same SNP and individual (e.g. AA for one pipeline and GG for the other) was negligible in almost all cases (from 0 to 0.098%). The number of SNPs obtained with reference genome (RG) approach was always lower than with both *de novo* building-loci pipelines (see Tables S3 and S5), however, the number of SNPs shared between RG and each of the *de novo* pipelines was similar (i.e. RG-STA and RG-ALT). The percentage of SNPs obtained with RG approach detected as well in both *de novo* pipelines was 47.7% (RG-STA/ RG) and 37.7% (RG-ALT/RG) in Manila clam and 73.0% (RG-STA/RG) and 80.9% (RG-ALT/RG) in brown trout, unlike the reverse way where the percentages of shared SNPs were lower due to the higher number of SNPs obtained with ALT: 28.8% (RG-STA/STA) and 11.4% (RGALT/ALT) for Manila clam with (Table S3) and 46.5% (RG-STA/STA) and 32.9% (RG-ALT/ALT) for brown trout (Table S5). Both *de novo* building-loci pipelines performed similarly when compared with reference genome (RG) approach genotyping (RG-STA and RG-ALT genotype comparisons; Table 2).

The population parameters evaluated (e.g. diversity levels, global F_{ST}) were roughly similar when using the different SNP panels in each species, showing higher differences in F_{IS} values (Table 1). The most notable differences would be found in the F_{IS} values and when comparing *de novo* and reference genome approaches in brown trout, especially regarding Hardy-Weinberg tests and related population parameters (i.e. F_{IS}, H_o *vs* H_e; Table S5). Here, unlike Manila clam, the proportion of SNPs with extremely low F_{IS} ($\leq$ - 0.5) greatly differed between both approaches. The structure patterns obtained using

STRUCTURE and DAPC analyses (see "Methods" section) were similar between both approaches (Figs. S5-S9). The highest global F_{ST} values among populations were found in brown trout and silver catfish, as expected, and no different interpretations among panels could be extracted. Some minor discrepancies in the number of suggestive outliers among panels were detected. All suggestive outliers detected showed a positive α-value suggesting divergent selection. The complete set of population parameters is provided in Supplementary Tables (Tables S3-S7).

Table 1. Mean (bold values) and standard deviation of population parameters for the final SNP panels using a *de novo* approach. Mean observed heterozygosity across loci and populations (H_o), mean expected heterozygosity across loci and populations (H_e), global fixation index (global F_{ST}), mean inbreeding coefficient across populations (F_{IS}), mean allelic richness across loci and populations (A_r), number of population structure units detected using STRUCTURE (STR groups) are shown. Y represents structure and N represents no structure. The complete information can be found in Supplementary Tables (Tables S3-S7).

		STA	ALT	COM	MER
M. clam	H_o (± SD)	**0.120** (0.014)	**0.103** (0.005)	**0.138** (0.017)	**0.108** (0.010)
	H_e (± SD)	**0.163** (0.005)	**0.135** (0.006)	**0.170** (0.008)	**0.140** (0.000)
	Global F_{ST}	0.006	0.003	0.004	0.005
	F_{IS} (± SD)	**0.237** (0.054)	**0.251** (0.034)	**0.195** (0.058)	**0.228** (0.044)
	A_r (± SD)	**1.698** (0.010)	**1.660** (0.024)	**1.713** (0.019)	**1.668** (0.015)
	STR (Groups)	N	N	N	N
C. cockle	H_o (± SD)	**0.145** (0.010)	**0.125** (0.006)	**0.150** (0.008)	**0.133** (0.005)
	H_e (±SD)	**0.157** (0.005)	**0.140** (0.000)	**0.160** (0.000)	**0.150** (0.000)
	Global F_{ST}	0.033	0.029	0.032	0.030
	F_{IS} (± SD)	**0.086** (0.028)	**0.120** (0.026)	**0.066** (0.036)	**0.114** (0.029)
	A_r (± SD)	**1.707** (0.026)	**1.683** (0.024)	**1.733** (0.024)	**1.690** (0.022)
	STR (Groups)	Y (3)	Y (3)	Y (3)	Y (3)

Table 1. (*Cont.*)

B. trout	H_o (± SD)	**0.243** (0.023)	**0.250** (0.035)	**0.200** (0.026)	**0.257** (0.029)
	H_e (± SD)	**0.190** (0.017)	**0.187** (0.021)	**0.170** (0.026)	**0.193** (0.023)
	Global F_{ST}	0.376	0.370	0.442	0.348
	F_{IS} (± SD)	**-0.269** (0.023)	**-0.336** (0.028)	**-0.179** (0.038)	**-0.333** (0.024)
	A_r (± SD)	**1.523** (0.041)	**1.520** (0.046)	**1.470** (0.044)	**1.533** (0.042)
	STR (Groups)	Y (2-3)	Y (2-3)	Y (2-3)	Y (2-3)
S. catfish	H_o (± SD)	**0.235** (0.049)	**0.235** (0.049)	**0.230** (0.057)	**0.235** (0.049)
	H_e (± SD)	**0.230** (0.056)	**0.230** (0.057)	**0.230** (0.057)	**0.235** (0.049)
	Global F_{ST}	0.452	0.453	0.465	0.451
	F_{IS} (± SD)	**-0.004** (0.032)	**-0.012** (0.036)	**-0.002** (0.024)	**-0.014** (0.038)
	A_r (± SD)	**1.690** (0.180)	**1.680** (0.170)	**1.685** (0.177)	**1.690** (0.170)
	STR (Groups)	Y (2)	Y (2)	Y (2)	Y (2)

Table 1. (*Cont.*)

S-s. catshark	H_o (± SD)	**0.545** (0.021)	**0.520** (0.000)	**0.460** (0.028)	**0.535** (0.007)
	H_e (± SD)	**0.355** (0.007)	**0.340** (0.000)	**0.325** (0.021)	**0.340** (0.000)
	Global F_{ST}	0.002	0.002	0.004	0.002
	F_{IS} (± SD)	**-0.528** (0.035)	**-0.541** (0.018)	**-0.406** (0.024)	**-0.544** (0.020)
	A_r (± SD)	**1.925** (0.021)	**1.905** (0.007)	**1.915** (0.021)	**1.915** (0.007)
	STR (Groups)	N	N	N	N

Table 2. Genotypic differences between shared SNPs from the different pipelines. Genotyping differences are presented as the relative frequency of total genotypes. COM panels were obtained with shared SNPs between STA and ALT panels (both *de novo* approach). RG-STA panels were obtained with shared SNPs between RG and STA (reference genome and *de novo* approach). RG-ALT panels were obtained with shared SNPs between RG and ALT (reference genome and *de novo* approach). Hom: Homozygous; MD: Missing data (i.e. missing genotype); Het: Heterozygous.

Species	SNPs Panel	Hom → Hom	Hom → MD	MD → Hom	Het → MD	MD → Het	Hom → Het	Het → Hom
	COM	0.00062	0.01500	0.01884	0.01324	0.00569	0.00349	0.01333
M. clam	RG-STA	0.00007	0.01759	0.00217	0.00567	0.00105	0.00336	0.00369
	RG-ALT	0.00017	0.01018	0.00450	0.01376	0.00300	0.00150	0.01243
C. cockle	COM	0.00004	0.00575	0.00557	0.01145	0.00273	0.00331	0.00475
	COM	0.00002	0.01285	0.01196	0.00952	0.01161	0.00407	0.00150
B. trout	RG-STA	0	0.01231	0.00708	0.00944	0.00021	0.00020	0.00744
	RG-ALT	0	0.00179	0.00564	0.00780	0.00040	0.00009	0.00060
S. catfish	COM	0.00019	0.00765	0.00520	0.00848	0.00382	0.00349	0.00547
S-s. catshark	COM	0.00098	0.04128	0.01311	0.05881	0.03522	0.01163	0.01393

4 POPULATION GENOMICS FOR CONSERVATION OF BROWN TROUT RESOURCES IN THE MEDITERRANEAN AND ATLANTIC SLOPES OF IBERIAN PENINSULA: STRUCTURE, SELECTION AND PATTERNS OF HYBRIDIZATION AND INTROGRESSION IN ATLANTIC AND MEDITERRANEAN RIVER DRAINAGES

4.1 INTRODUCTION

Brown trout is an important species in freshwater ecosystems across all its natural distribution, including the Iberian Peninsula (*Observatorio Español de Acuicultura*, OESA 2018) from ecological, recreational, and economic points of view. In Spain there is a high freshwater angling activity, with 490,868 fishing licenses issued in 2018. Within Galicia, there were around 35,000 fishing licenses in 2019 (*Consellería de Medio Ambiente, Territorio e Vivenda* 2019). In this Autonomous Community, fish farms for freshwater species are basically dedicated to rainbow trout (*Oncorhynchus mykiss*) culture, with more than 20 facilities (ARDÁN-*Informe económico y de competitividad* 2018), although four hatchery facilities are devoted to produce brown trout eggs or fry for restocking. As outlined before, *S. trutta* has a high socioeconomic interest due its human-mediated

exploitation. According to ichthyoarchaeological analyses, brown trout consumption was practiced since the Neolithic in human communities of the Sierra de Atapuerca, where brown trout is common nowadays (Blanco-Lapaz and Vergès 2016). The existence of brown trout as a species occurred around 10-14 Ma (Lecaudey *et al.* 2018) and the divergence in the *S.trutta* complex took place during the Pliocene, around 2.5-5 Ma (Crête-Lafrenière *et al.* 2012). Since then, brown trout populations occupied the Iberian rivers through the recurrent glaciation events associated to the Pleistocene, when the Iberian Peninsula was an important refugia for many species due to its warmer climatic conditions, which later expanded northwards from this southern European refugium (Weiss 2010).

The analysis of mtDNA sequence in the last three decades has allowed identifying four Iberian native lineages: Duero (DU), Atlantic (AT), Adriatic (AD) and Mediterranean (ME). The first one would be an endemism of the Iberian Peninsula (Vera *et al.* 2010), located in the inner sections of the Duero (Bouza *et al.* 2008; Hermida *et al.* 2009; Martínez *et al.* 2007; Vera *et al.* 2010) and Miño-Sil basins (Vera *et al.* 2015; Vilas *et al.* 2010). These mitochondrial lineages are characterized by different haplotypes (Cortey *et al.* 2009; Vera *et al.* 2010). Moreover, these four matriarchal lineages are geographical distributed. DU and AT lineages are distributed in the western slope of Iberian Peninsula, while AD and ME lineages in the Mediterranean one. DU and AT brown trout are distributed in the Miño and Duero basins following a spatial segregation in parapatry with a not well-defined hybrid zone between them (Bouza *et al.* 2001; Martínez *et al.* 2007). This hybrid zone would be the consequence of secondary contacts (Bouza *et al.* 2001) between the endemic DU lineage, ancestral in the Iberian Peninsula, and the nowadays more abundant AT arrived in later expansions from the north (García-Marín *et al.* 2018). On the other hand, AD and ME lineages are sympatrically distributed, although following a mosaic patchiness in the

Mediterranean slope (Cortey *et al.* 2004; Sanz *et al.* 2002; Vera *et al.* 2013). Resident and migratory (sea trout) forms of brown trout are present in Galicia, since the Miño outlets at parallel 42º N, the distribution limit of this migratory form (Bouza *et al.* 1999).

In addition to the phylogeographic variation of brown trout in the Iberian Peninsula previously described, there is a foreign genetic lineage due to releases of hatchery brown trout using a stock from Central European origin commonly used for restocking practices throughout the European continent. During decades, millions of fertile brown trout individuals have been released as eggs or fry into Spanish rivers to satisfy angling demand and counterbalance the decline of wild populations. The hatchery stock used was founded with the aforementioned strain belonging to an AT clade different than the wild AT lineage living in Iberian Peninsula (Cortey *et al.* 2009; Machordom *et al.* 2000). The genetic consequences of this restocking activity have been studied with allozyme markers by different authors, who detected different degrees of introgression throughout the Iberian Peninsula (Almodóvar *et al.* 2006; Martínez *et al.* 1993; Morán *et al.* 1991). The genetic introgression detected in Galician rivers using the diagnostic marker *LDH-C**90 was low, suggesting very low viability of hatchery individuals in Galician rivers (Arias *et al.* 1995; Martínez *et al.* 1993). Nevertheless, a much higher impact of restocking was detected in non-flowing waters (lagoons and reservoires; Martínez *et al.* 1993). The lower impact of restocking in the Atlantic Drainage has been related to the stable hydrological conditions of this region regarding Mediterranean Drainage, where the impact of restocking is higher (Almodóvar *et al.* 2006; Vera *et al.* 2013).

For conservation purposes the first is to define what to conserve. There are two main concepts: Management Units (MUs; i.e. populations within species that are genetically distinct enough to require separate management) and Evolutionary Significant Units

(ESUs), "whose divergence can be measured or evaluated by putting differential emphasis on the role of evolutionary forces at varied temporal scales" (Casacci *et al.* 2014). For the latter there are different definitions from the original by Ryder (1986) that can be inspected in Casacci *et al.* (2014; see Table 1), but all have in common the sentence in quotation marks outlined above. MUs and ESUs are not synonymous concepts (see Moritz 1994). To define specific management units, population genetic structure approaches are necessary. Furthermore, in species such as brown trout characterized by high population structure due to different factors (e.g. habitat patchiness; Ferguson 1989), defining MUs is a matter of controversy, since populations separated by a few kilometres show significant differentiation (Bouza *et al.* 1999; Carlsson and Nilsson 2001). This issue was the goal of previous studies with allozymes (Bouza *et al.* 1999, 2001; Martínez *et al.* 1993), microsatellites (Vilas *et al.* 2010) and mtDNA markers (Bouza *et al.* 2008). Two different mitochondrial lineages of brown trout co-occur in Galicia parapatrically segregated, DU and AT, such as in Duero Basin (Bouza *et al.* 2001; Vera *et al.* 2010). This distribution has been hypothesized as the consequence of a secondary contact between those two divergent lineages, being particularly remarkable in both basins, where putative hybrid zones would exist (Bouza *et al*. 2001; Martínez *et al.* 2007; Vilas *et al.* 2010). Within DU lineage, an endemism of the Iberian Peninsula (Machordom *et al.* 2000), there are different haplotypes with significant divergence, advising against treating this lineage as a single ESU (Vera *et al.* 2015). Within Mediterranean Slope, in Catalonia there is the figure of conservation of 'genetic refuges', where releases from hatchery stock are banned (Araguas *et al.* 2017). In other basins, it has been proposed to define additional refuges to protect remnant native brown trout lineages (AD and ME; Vera *et al.* 2013).

Over the last century, specific legislation has been developed in Spain concerning the regulation of angling, environmental protection, and the protection of biodiversity. Different laws were developed from the most specific, as Salmon Protection Law (1912), to those with larger spectrum such as Inland Fisheries Law (1907, 1929, 1942). These laws reflected the need for the regularization of the activity with a conservationist perspective. Also, within the European Union, the Council Directive 78/659/EEC of 18 July 1978 reflected the need for good quality fresh water to make life possible for the different fishes, with mention to salmonid and cyprinid water bodies. This Directive was replaced by Directive 2000/60/EC (*Directiva Marco del Agua*; DMA), designed to protect and improve freshwater quality, setting standards for defining water quality based on its "ecological status", using among other biological quality indicators. Over the years, the preservation of native genetic biodiversity has been incorporated into increasingly advanced laws as research has progressed (Box 1). By the same way, regulatory and conservationist legislation has been implemented in different autonomous communities (see *Ley 22/2009 de 23 de diciembre, de ordenación sostenible de la pesca en aguas continentales de Cataluña*; Box 1). Regarding freshwater angling in Galicia, during almost thirty years, the *Ley 7/1992 del 24 de julio, de pesca fluvial de Galicia* have been active. Recently, a new law was approved at the Galician Parliament (i.e. *Ley 2/2021, de 8 de enero, de pesca continental de Galicia*). There are some highlights respect to previous one (Box 2).

Box 1. Selection of legislative highlights at different Public Administration levels.

Artículo 45 de la Constitución Española (1978).

"1. Todos tienen el derecho a disfrutar de un medio ambiente adecuado para el desarrollo de la persona, así como el deber de conservarlo.

2. Los poderes públicos velarán por la utilización racional de todos los recursos naturales, con el fin de proteger y mejorar la calidad de la vida y defender y restaurar el medio ambiente, apoyándose en la indispensable solidaridad colectiva.

3. Para quienes violen lo dispuesto en el apartado anterior, en los términos que la ley fije se establecerán sanciones penales o, en su caso, administrativas, así como la obligación de reparar el daño causado."

Directiva del Consejo 78/659/EEC de 18 de julio de 1978, relativa a la calidad de las aguas dulces que deben protegerse o mejorarse para mantener la vida de los peces.

Artículo 1.1: "La presente Directiva trata de la calidad de las aguas continentales y se aplicará a las aguas que requieren protección o mejora para ser aptas para la vida de los peces, declaradas como tales por los Estados miembros."

Ley 42/2007, de 13 de diciembre, del Patrimonio Natural y de la Biodiversidad.

Artículo 67: "El Inventario Español de Caza y Pesca, dependiente del Ministerio de Medio Ambiente, mantendrá la información más completa de las poblaciones, capturas y evolución genética de las especies cuya caza o pesca estén autorizadas, con especial atención a las especies migradoras."

Ley 22/2009 de 23 de diciembre, de ordenación sostenible de la pesca en aguas continentales de Cataluña.

Artículo 29.2: "Las zonas de pesca controlada intensiva deben ubicarse en cursos, tramos de cursos o masas de agua transformados artificialmente, en especial los embalses, y fuera de las aguas de reserva genética, para evitar la degradación biológica y, en especial, genética de las poblaciones de especies autóctonas."

Artículo 47.3: "No pueden emplazarse nuevos centros industriales de producción de fauna en aguas continentales en derivaciones de tramos de cursos de agua que hayan sido declarados refugios de pesca ni en aguas de reserva genética."

Box 2. Selection of highlights from *Ley 2/2021, de 8 de enero, de pesca continental de Galicia*.

Artículo 10 (Investigación en materia de pesca continental): "La consejería competente en materia de pesca continental impulsará la mejora del conocimiento sobre la etología, biología y dinámica poblacional de las especies de la fauna acuática, en especial de las pescables. Asimismo, impulsará el conocimiento genético de las poblaciones ictícolas y del impacto de las especies exóticas invasoras sobre el ecosistema acuático, y la mejora de los métodos de gestión de la pesca continental."

Artículo 61.4 (Sueltas): "Las sueltas se realizarán con especies autóctonas y con ejemplares nacidos en libertad o procedentes de centros ictiogénicos dependientes de la consejería competente en materia de pesca continental y obtenidos de reproductores capturados en la misma cuenca hidrográfica en la que se va a realizar la suelta o, en su defecto, con ecotipos de la mayor similitud genética posible."

Artículo 62.2 (Repoblaciones piscícolas): "Las repoblaciones piscícolas se realizarán con especies autóctonas y con ejemplares nacidos en libertad o procedentes de centros ictiogénicos dependientes de la consejería competente en materia de pesca continental y obtenidos de reproductores capturados en la misma cuenca hidrográfica en la que se va a realizar la repoblación o, en su defecto, con ecotipos de la mayor similitud genética posible."

Artículo 62.4 (Repoblaciones piscícolas): "No podrán repoblarse aquellos tramos de agua en los que habiten poblaciones piscícolas de interés por sus peculiaridades biológicas o genéticas, así como aquellos tramos de agua en los que exista algún régimen de protección especial, salvo por razones de defensa de las poblaciones, debidamente justificadas."

Artículo 63.1 (Centros ictiogénicos): "Se declaran de interés general los centros ictiogénicos para el fomento de la recuperación y conservación de las poblaciones piscícolas salvajes y del medio en el que se desarrollan."

The specific aims of this chapter were: (1) Identification of farmed and wild samples and development of diagnostic SNP panels; (2) to analyse patterns of human-mediated introgression across the genome; (3) to analyse the pattern of genetic diversity and structure through populations, and (4) to find traces of natural selection.

4.2 MATERIAL AND METHODS

4.2.1 Sampling

Thirteen populations of brown trout (*Salmo trutta*) from Miño and Duero River basins draining into the Atlantic Ocean, and Ter Basin into the Mediterranean Sea from the Iberian Peninsula were sampled (Table 3). All populations from Miño and Duero basins were collected in tributaries (Fig. 11). Populations belonging to the Atlantic Slope included representatives of the endemic DU lineage (LE, P2 and P3), AT lineage (VI, AG1, and BL), and presumably hybrid populations (FE, CH, CE, and OM) according to previous isozyme, mtDNA and microsatellite data (Bouza *et al.* 1999, 2001, 2008; Martínez *et al.* 2007; Vilas *et al.* 2010). Mediterranean populations belonged to some of tributaries of Ter Basin (Núria and Freser), including two temporal replicates from Núria and Ter rivers (NU04-NU14 and TE04-TE14, respectively; Table 3, Fig. 12). Different restocking incidence with the Central European hatchery stock had been reported previously on the Ter Watershed (Araguas *et al.* 2017). Two representative samples from the Bagà hatchery stock used for restocking in Catalonia were included in our analysis. Despite there are different hatchery stocks across the Iberian Peninsula they are quite homogeneous due to their recent common origin (García-Marín *et al.* 1991), and furthermore, even across Europe (Bohling *et al.* 2016). So, this hatchery was used as reference to detect individuals of hatchery ancestry in wild populations in the samples analysed in this study. Duero Basin is interrupted by big hydroelectric dams built mainly in the 50'-60's of XX century, while the Mediterranean populations mostly by dams lower than three metres from similar construction dates. A total number of 299 individuals captured by electrofishing were analysed, mainly 0+, 1+ and 2+ classes, and included the four native mitochondrial lineages previously identified in the Iberian Peninsula (i.e. AD, AT, DU and ME; Table 3). Furthermore, the samples studied included the two distribution patterns

known in the studied area: (1) parapatry or spatial segregation with a hybrid contact zone between AT and DU lineages in Miño and Duero basins; and (2) mosaic-sympatry between AD and ME lineages in Mediterranean Drainage.

Table 3. Characteristics of the brown trout (*S. trutta*) samples from Iberian Peninsula analysed.

Origin	*Code*	*No. individuals*	*mtDNA lineage*
Miño-Sil Basin		***59***	
Viñao River (2003)	VI	16	(AT)
Ferreira River (2003)	FE	13	(AT/DU)
Chamoso River (2003)	CH	14	(AT/DU)
Lea River (2003)	LE	16	(DU)
Duero Basin		***119***	
Águeda River (2002)	AG1	20	(AT)
Porto do Rei Búbal River (2002)	BL	19	(AT)
Cega River (2002)	CE	20	(AT/DU)
Omaña River (2002)	OM	20	(DU)
Pisuerga River 2 (2002)	P2	20	(DU)
Pisuerga River 3 (2002)	P3	20	(DU)
Hatchery		***39***	
Hatchery release individuals (2014)	BA14	19	(AT)
Hatchery spawners (2014)	S	20	(AT)
Catalonia river basins		***82***	
Núria River (2004)	NU04	16	(AD/ME)
Núria River (2014)	NU14	16	(AD/ME)
Queralbs, in Freser River (2014)	QB14	18	(AD/ME)
Ter River (2004)	TE04	14	(AD/ME)
Ter River (2014)	TE14	18	(AD/ME)

mtDNA lineage: Atlantic (AT), Duero (DU), Adriatic (AD) and Mediterranean (ME)

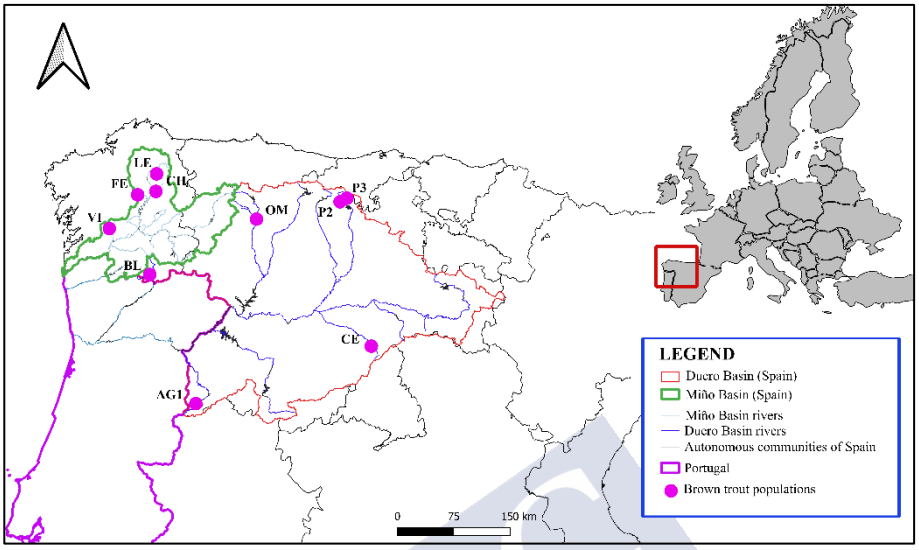

Figure 11. Location of brown trout (*S. trutta*) samples in Miño and Duero basins. Projection ETRS89 UTM zone 30 N (EPSG:25830). Map made with QGIS 3.10.

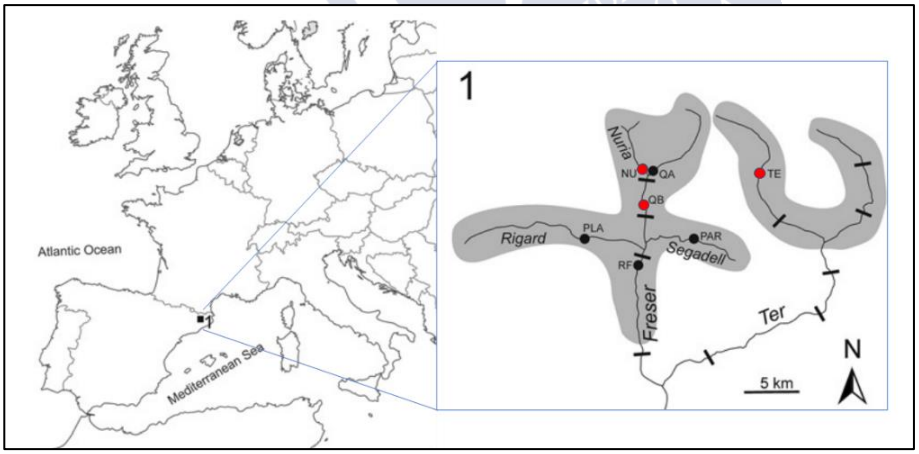

Figure 12. Location samples and dams in the Ter Basin from the Mediterranean Drainage. Red dots and shaded areas represent sampling points and genetic refuges areas, respectively. Modified from Araguas *et al.* (2017).

4.2.2 Identification and genotyping of SNPs: defining a reference panel for all populations studied

DNA extraction and 2b-RAD libraries preparation followed the protocol described by Wang *et al.* (2012) with slight modifications. AlfI was the IIB restriction enzyme (RE) used for digestion of genomic DNA to construct the libraries to be sequenced on a NextSeq500 Illumina sequencing platform following a 50 bp single-end chemistry. The number of recognition sites in *S. trutta* genome was obtained *in silico* using the last version of ExtractSites.pl script (https://github.com/Eli-Meyer/2brad_utilities) taking as reference the brown trout genome (NCBI accession GCA_901001165.1).

After raw data demultiplexing, several filtering criteria were applied: (1) all reads were trimmed to the expected length (i.e. 36 nucleotides), (2) filtered by the RE recognition site presence using own Perl scripts, and (3) cleaned with process_radtags (module belonging to Stacks 2.41) to remove reads with at least one uncalled nucleotide or nine or more consecutive nucleotides with average Phred quality scores lower than 30 (base call accuracy of 99.9%; −w 0.25, −s 30, −c, −q). Filtered reads were aligned to the brown trout genome with Bowtie 1.3.0 (Langmead *et al.* 2009) using the -v alignment mode (v = number of allowed mismatches; -v 3). Reads with more than one alignment with the same quality were discarded (--best --strata -m 1). Since individuals with low read numbers with respect to the rest of the sample might act as a burden to obtain a robust SNP panel, individuals with less than 900,000 aligned reads (Q_1 aligned reads – IQR) were removed. Stacks input alignments were oriented with the same sense using own perl script and sorted by chromosome using Samtools 1.10 (Li *et al.* 2009).

The Stacks pipeline for further analyses consisted in two modules: (1) gstacks, which genotypes SNPs identified per locus in each individual; and (2) populations, which obtains a variety of standard

output formats for further population genetics analyses (e.g. Genepop format).

The initial panel of SNPs obtained was constituted mainly by nuclear SNPs and a few mtDNA SNPs. This panel was analysed to detect the same SNPs from adjacent RAD-loci, caused by the proximity of AlfI recognition sites. Next, the raw SNP panel was filtered to retain a consistent set of markers and alleles represented across the individuals genotyped according to the following criteria: (1) coverage ≥ 6 reads per locus and individual (MinDP ≥ 6x), (2) least frequent allele score ≥ 3 alleles in the whole sample (MAC ≥ 3), (3) genotyped in at least in 60% of the individuals in each population, (4) conformance to Hardy-Weinberg (HW) expectations (p-value > 0.05; Fisher's method) in more than half of the populations analysed, (5) ≤ 3 SNPs per RAD-locus, (6) when several SNPs occurred in a single RAD-locus, the most polymorphic was retained to avoid redundant information, and (7) consistently homologous to a single position in the main units of the brown trout genome assembly (i.e. chromosomes).

4.2.3 Identification of farmed and wild samples

Because restocking practices have been carried out during decades in the studied regions of Iberian Peninsula (Araguas *et al.* 2017; Bouza *et al.* 1999, 2001; Martínez *et al.* 1993; Vilas *et al.* 2010), individuals of hatchery ancestry were identified and removed for genetic analyses of wild populations. A single diagnostic locus (*LDH-C**), fixed for the *100 and *90 alleles in the wild and hatchery stocks, respectively, has been used to identify individuals of hatchery ancestry in the Iberian Peninsula (Morán *et al.* 1991). However, the availability of a genomic screening in our study offered the opportunity to develop a more sophisticated and powerful tool for this purpose. As a first step, SNPs were selected according to differentiation between hatchery and wild Iberian populations here studied (F_{ST} > 0.95). The Bayesian clustering method implemented in STRUCTURE 2.3.4 (Pritchard *et al.* 2000)

using R package ParallelStructure 1.0 (Besnier and Glover, 2013) was used to ascertain the proportion of hatchery ancestry (*qH*) in each individual from the wild using BA14 (Bagà) as a hatchery reference population. As indicated above, the BA14 stock was founded from the original Centro-European stock imported for restocking in Spain and shows a very similar genetic composition to all others used for this purpose in the Iberian Peninsula (Almodóvar *et al.* 2006). An admixture model with two populations was performed (hatchery and native; K = 2), where hatchery individuals were forced to be non-admixed (i.e. POPFLAG = 1), following an incomplete baseline method frequently used (Sanz *et al.* 2009; Vera *et al.* 2013). Due to the high differentiation reported between Iberian native populations and the hatchery stocks, a model of independent allele frequencies was used (see Sanz *et al.* 2009). For STRUCTURE analyses burn-in of 100,000 iterations, 200,000 Markov Chain Monte-Carlo steps (MCMC) were applied. Ten independent replicate runs were used to limit the influence of stochasticity and increase the precision of estimations following recommendations by Gilbert *et al.* (2012). SNPs from the whole panel that resulted to be singletons when using only two populations (hatchery and wild) were removed for analyses as recommended by Linck and Battey (2019). The Genepop subsets were obtained with genepopedit 1.0 R package (Stanley *et al.* 2017) and converted into STRUCTURE files using PGDSpider 2.1.1.5 software (Lischer and Excoffier 2012). A *qW* (W = wild cluster) threshold was established using the lowest value of self-membership coefficients (*q*) obtained in a panmictic population available in our study (BA14 hatchery stock) using POPINFO = 0. Accordingly, individuals with *qW* values < 0.95 were considered of farmed ancestry and removed for further analyses of wild population structure in Iberian Peninsula. All individuals were classified in three categories: (1) wild individuals (*qW* > 0.95), (2) introgressed hatchery ancestry individuals (*qW* between 0.95 and 0.05), and (3) hatchery individuals (*qW* < 0.05); anyway, the last two classes

were collapsed in a single class to gather all individuals with any hatchery ancestry component detected in the wild. Additionally, the same analysis was performed with three different subsets of discriminatory SNPs with the purpose of setting up a new cheap and powerful molecular tool to identify hatchery ancestry individuals in the wild Iberian populations. For this, the same STRUCTURE analysis was run with three different subsets of physical unlinked diagnostic SNPs: $F_{ST} > 0.95$, $F_{ST} > 0.99$ and $F_{ST} = 1$ (i.e. fully diagnostic SNPs with fixed allelic variants between wild and hatchery populations) and results were compared with those obtained with the whole genomic information outlined before. These F_{ST} values were obtained between hatchery and Iberian populations conformed exclusively by wild individuals ($qW >$ 0.95). The proportion of correct classification of hatchery ancestry individuals with each subset of SNPs in every population analysed regarding those detected with the whole genomic information was used to establish the performance of each SNP dataset. This information could be applied in the future to: (1) identify hatchery ancestry individuals using cheaper and robust genotyping techniques such as SNaPshot® (Applied Biosystems) or MassARRAY (Sequenom); (2) bioinformatic analysis, saving time or when large hardware resources are not available (e.g. supercomputing centre).

The inference of local ancestry was performed with LAMP 2.5 software in the LAMPANC mode (Sankaraman *et al.* 2008), which relies in a naive Bayes classifier and a clustering algorithm (Iterated Conditional Modes; ICMs) to find an optimal classification of each individual in terms of the likelihood. LAMP estimates the most probable ancestry at a site using a number of SNP sliding windows with their corresponding ancestry estimates that will overlap with the same SNP. Majority vote over windows will be used to call SNP's ancestry (Sankaraman *et al.* 2008). The affected population with the largest sample size of admixed individuals was QB14 (78%, N = 14; see

Results). Then, local ancestry inference was exclusively performed with QB14 population. Two ancestral genetic pools were used: BA14 and TE14, representing the hatchery and wild references, respectively. The program was run separately for SNPs belonging to each of the 40 *S. trutta* chromosomes. Only SNPs with $F_{ST} > 0.10$ between the two reference pools were used. The recombination rate used was 0.88 cM/Mb. This estimation was obtained by Leitwein *et al.* (2017) for the whole *S.trutta* genome. The number of generations since the beginning of mixing was set to ten, assuming restocking in this region since 80s (see Table 1 in Araguas *et al.* 2004). The r^2 cut-off was set to 0.30 since the small sample size may lead to linkage disequilibrium (default value: 0.10). Above this threshold two SNPs would be considered non independent and only one would be retained for the ancestry estimation. The admixture fraction (alpha) was estimated with STRUCTURE 2.3.4. Local ancestry along genome was plotted using RIdeogram 0.2.2 R package (Hao *et al.* 2020).

4.2.4 Genetic diversity

Observed and expected heterozygosities (H_o and H_e) and allelic richness with the rarefaction method were used to estimate genetic diversity per population with DiveRsity R package 1.9 (Keenan *et al.* 2013) using the 'basicStats' function. The degree and sense of the deviation from panmixia was estimated with the intrapopulation fixation index (F_{IS}; Wright 1951). Bias corrected 95% Confidence Intervals (CI) were obtained after 1,000 bootstraps iterations with 'divBasic' function. Conformance to Hardy-Weinberg equilibrium (HWE) was tested with R package Genepop 1.1.7 (based on the Genepop 4.7.5 version; Rousset 2008). To obtain *p*-values the complete enumeration method was used (Louis and Dempster 1987). Global *p*-values were obtained with Fisher's method. Number of private alleles per populations was obtained with PopGenReport 3.0 R package (Adamack and Gruber 2014) using the 'allele.dist' function.

4.2.5 Genetic differentiation and structure

Global and pairwise coefficients of population differentiation (F_{ST} values; see Weir and Cockerman 1984) were estimated using different hierarchical criteria. Pairwise F_{ST} between populations was calculated using StaMPP R package 1.6 (Pembleton *et al.* 2013) with the 'stamppFst' function. For this, 10,000 bootstrap replicates across loci to generate 95% confidence intervals and *p*-values regarding the null hypothesis ($F_{ST} = 0$) were used. Global F_{ST} for the whole dataset and for each region considered was calculated using R package Genepop 1.1.7 with the 'Fst' function.

Four different statistical methods were applied to investigate genetic structure in wild populations using all SNPs, including non-parametric approaches (e.g. Discriminant Analysis of Principal Components, DAPCs) approaches, as recommended by Linck and Battey (2019): (1) Analysis of the MOlecular VAriance (AMOVA; Excoffier *et al.* 1992), (2) bayesian clustering method with STRUCTURE, (3) DAPCs with cross validation and (4) DAPCs with the number of Principal Components (PCs) comprehending ≥ 90% of the variance (e.g. Ríos *et al.* 2020; Vera *et al.* 2018). Further STRUCTURE analyses were performed for wild hybrid zones previously identified (Bouza *et al.* 2001; Martínez *et al.* 2007; Vilas *et al.* 2010) with a selection of AIMs (Ancestry Informative Markers) with $F_{ST} > 0.50$ between reference populations.

AMOVAs were performed with Arlequin 3.5.2.2 (Excoffier and Lischer 2010), computing F-statistics derived from different hierarchical partitions. Statistical significance of F-statistics for each scenario was tested with 10,000 permutations. With this approach, the best grouping scenario maximizes F_{CT} value (relative component of diversity among groups) by reducing F_{SC} one (*idem* among populations within groups). Nine *a priori* grouping scenarios were independently analysed: (1) Atlantic *vs* Mediterranean slope populations; (2)

populations grouped according to their basin (i.e. Miño, Duero and Ter basins); (3) Miño and Duero basin populations; (4) across temporal replicates and localities in Mediterranean Slope; (5a) Duero Basin populations grouped according to STRUCTURE cluster composition (see Results section); (5b) populations grouped according to Bouza *et al.* (2001); (5c) populations grouped according to Martínez *et al.* (2007); (6a) Miño basin populations grouped according to STRUCTURE determined cluster composition and (6b) populations grouped according to Bouza *et al.* (2008) and Vilas *et al.* (2010).

For STRUCTURE and DAPC analyses, samples within the temporal range 2002-2004 were used. STRUCTURE analyses were run with the same admixture model and the same length in burn-in and MCMC described above. Non *a priori* population information was used (POPINFO = 0). When only populations from one basin were included, correlated allele frequency models were used, nevertheless, when populations from different basins were included independent allele frequency models were applied due to the high differentiation reported between populations belonging to different basins. StructureSelector web based software (Li and Liu 2018) was used to obtain K estimators and CLUMPAK outputs (Kopelman *et al.* 2015). Three K estimators were used to identify the most likely number of clusters: the deltaK *ad hoc* estimator (Evanno 2005), Mean LnP(K) (Pritchard *et al.* 2000) and MedMeaK (Puechmaille 2016). CLUMPAK output files rendered STRUCTURE bar plots illustrating membership of individuals to inferred genomic clusters.

DAPC 2-step approach: A Principal Component Analysis (PCA) from the matrix of the genotypes is performed and then, a selected number of Principal Components (PCs) instead of the original SNP genotypes is used as input for the linear discriminant analysis (LDA). Initially, the 'find.cluster' function, implemented in the software Adegenet 2.1 (Jombart and Ahmed 2011), working with all PCs was

applied to determine the best supported number of genetic clusters using the bayesian information criterion (BIC). The 'find.cluster' function runs successive K-means clustering with increasing number of clusters (K) and provides a BIC value for each evaluated K-value, where the lowest BIC is the "optimal" number of clusters. The maximum number of clusters assayed was the double the number of populations and 100,000 iteration per run and 100 starting centroids were used. The selection of the optimal number of PCs to be further used in the LDA was done via cross-validation method where the data are split into: a training set (90% of the data) and a validation set (10% of the data). Cross validation was carried out in two steps ('xvalDapc' function): (1) a maximum number of 300 PCs were tested with 100 replicates; (2) with the results obtained, a second cross-validation was run by specifying a narrow range of PCs with 1,000 replicates. The best number of PCs retained was associated with the lowest Root Mean Square Error (RMSE). Additionally, DAPCs retaining at least 90% of the cumulative variation of the data were performed. The resultant clusters were represented in a 2D-scatterplot using the best linear components of DAPC. After this, the software GENECLASS2 (Piry *et al.* 2004) was used to ascertain the presence of first-generation immigrants (F0) at populations. Since plausibly not all immigrant recipient populations are sampled, the L-home option (L_h test statistic; Paetkau *et al.* 2004) of GENECLASS2 was used (L_h; the likelihood of drawing that individual's genotype from the population in which it was samples). The simulation algorithm of Paetkau *et al.* (2004) with 1,000 simulated individuals was applied.

4.2.6 Effective population size estimation

Contemporary effective population size (N_e) was estimated for each population using NeEstimator 2.1 software (Do *et al.* 2014) with the Linkage Disequilibrium (LDN_e) method under a random mating model (Waples 2006). In the two Mediterranean locations where,

temporal replicates were available (Núria and Ter), three different formulations of the temporal method (Jorde and Ryman 2007; Nei and Tajima 1981; Pollak 1983) and generation sets were used. The 95% confidence intervals were determined using the non-parametric jack-knife method, more recommendable when the number of loci is large (> 100; see NeEstimator Help file). To prevent potential biases introduced by low frequency alleles, singleton alleles were removed in all analyses to ensure a minimum allele frequency (MAF) > 1/2N. Due to linkage disequilibrium can be generated by the sampling process itself (England *et al.* 2006) and the small sample size (N~15), when LDN_e method was applied, seven minimum allele frequency thresholds were used (between 0.10 and 0.40). Accordingly, the effect of different MAF thresholds over N_e estimates was evaluated as suggested by Marandel *et al.* (2020). N_e estimates were considered confident when this subsampling SNP method according to MAF reached a plateau of N_e estimates. Finally, the impact of physical linkage among markers was evaluated with N_e estimates from comparisons between SNPs placed in different chromosomes to calculate r^2.

4.2.7 Adaptative variation

SNPs in genomic regions under selection (outlier dataset) were identified by both Arlequin 3.5.2.2 and BayeScan 2.1. Different subsets of the total dataset were set up by removing SNPs that were monomorphic or singletons in sample subsets (e.g. Miño Basin) in the same way as in the structure analyses. Different statistical approaches were applied to identify a confident set of outlier loci as recommended Narum and Hess (2011). The BayeScan procedure removes the effects shared by all loci (beta), influenced by genetic-drift, from locus-specific effects (alpha), potentially driven by selection. Departure from neutrality at a given locus is assumed when alpha is significantly different from 0. The values of alpha can be informative about the type of selection (i.e. positive values suggest divergent selection and

negative values suggest balancing selection). For these analyses, very small sample size can be used assuming the lower power of the tests performed, but with no particular risk on bias. The BayeScan analyses were carried out for 20 pilot runs, 5,000 iterations, 100,000 burn-in steps and -pr_odds flag 10 (i.e. the odds for a neutral evolution model were 10 times higher than a model including selection). Loci with *q*-values (False Discovery Rate, FDR) < 0.05 were considered significant outliers. In Arlequin, two models for detection of loci under selection were implemented: (1) the finite island model and (2) the hierarchical island model (as defined by Slatkin and Voelm 1991) to avoid a large fraction of false positives when populations share a recent history or belong to a hierarchically subdivided population (Excoffier *et al.* 2009). For the finite island model, Arlequin was set up for testing 100,000 simulations with 1,000 demes, and when using a hierarchical finite island model 100,000 simulations with 1,000 simulated demes and 100 groups were set up. For detecting outlier loci, F_{ST} was used in both models, as recommended by the Arlequin manual. Loci with *p*-value < 0.01 were considered as significant outliers considering the tendency of this program to false positives (Narum and Hess 2011). For the hierarchical finite island model different grouping with significant F_{CT} in AMOVA analyses were used.

Outliers were analysed (1) for all samples belonging to the temporal range 2002-2004; (2) among populations belonging to Atlantic Slope (i.e. Miño and Duero basins); (3) among populations from Duero Basin; (4) among populations from Miño Basin. The obtained outliers were classified into two categories: (1) suggestive outliers, those detected in any of the methods applied; (2) consistent outliers, those detected with all methods. Non-synonymous substitutions due to allelic variants at SNPs located in exons were evaluated. Open reading frames (ORFs) were checked with ORFfinder (NCBI) and BioEdit 7.2.5 (Hall 1999) using coding DNA sequences

(CDSs) from the brown trout database in Ensembl (https://www.ensembl.org/Salmo_trutta/). Gene mining to identify candidate genes under selection was performed on those genomic regions where two or more consistent outlier loci were placed in the same chromosome region within a range < 500 kb. Gene annotation was performed with the Ensembl database 103. GO terms were obtained with Blast2GO (Götz *et al.* 2008), module of the software OmicsBox 1.4.12 (https://www.biobam.com/omicsbox-update-1-4/).

4.3 RESULTS

The number of potential RAD-loci detected *in silico* in the brown trout genome was 557,761, representing ~0.85% of the total genome assembly size. Among them, five RAD-loci were in the mitochondrial genome. The number of 2bRAD-loci per chromosome was highly correlated with chromosome length (R^2 = 0.98, *p*-value = 0.000). On average one 2bRAD-locus occurred every 4,000 nucleotides.

A total of 2,054,115,335 raw reads from 299 individuals were produced on the NextSeq500 sequencing platform, averaging 6,869,950 reads per individual. After filtering steps 1,410,823,259 reads were retained (68.7%). The most astringent filtering step was RE recognition site presence (~20% removed reads). Out of those reads, 610,183,196 aligned against the brown trout reference genome with -v 3 (43.2%). Sixteen individuals with less than 0.9 M aligned reads were discarded (Table 4).

Table 4. Number of individuals per sample used for subsequent analyses (N = 283), classified into natural basins or hatchery.

Origin	*Code*	*No. individuals*	*No. final individuals*
Miño-Sil Basin		***59***	***56***
Viñao River	VI	16	15
Ferreira River	FE	13	13
Chamoso River	CH	14	13
Lea River	LE	16	15
Duero Basin		***119***	***106***
Águeda River	AG1	20	16
Porto do Rei Búbal River	BL	19	16
Cega River	CE	20	19
Omaña River	OM	20	20
Pisuerga River 2	P2	20	18
Pisuerga River 3	P3	20	17
Hatchery		***39***	***39***
Hatchery release individual	BA14	19	19
Hatchery spawners	R	20	20
Catalonia river basins		***82***	***82***
Núria River	NU04	16	16
Núria River	NU14	16	16
Queralbs, in Freser River	QB14	18	18
Ter River	TE04	14	14
Ter River	TE14	18	18

A total of 361,129 RAD-loci were built by Stacks 2.41, comprising 606,037,805 aligned reads and representing an average 9.0x coverage per locus and individual. After Stacks 2.41 parsing, 191,406 SNPs were retained. Among these, 1,150 SNPs (0.6%) were within overlapping RAD-loci, being discarded for further analyses. Among the next filtering steps, the main SNP dropping was due to population representation ≥ 60% (66.7%; Table 5) and then the MAC ≥ 3 filtering step (41.1%).

Table 5. Filtering steps for the final SNP panel.

Filtering step	Number of SNPs
Stacks 2.41 output	191,406 (6 mitochondrial SNPs)
MinDP ≥ 6x	191,406
MAC ≥ 3	112,652
SNPs represented 60%/pop	37,552
HW (p-value > 0.05)	34,583
No overlapping SNPs	34,548
RAD-loci with ≤ 3 biallelic SNPs	32,981
1 SNP/RAD-locus selected (more pol.)	25,247 (3 mitochondrial SNPs)
Well-located SNPs	24,830

Three mitochondrial SNPs were conserved after filtering steps, all of them located within coding regions. None of the final RAD-loci, either nuclear- or mtDNA-linked, had uncalled nucleotides in the reference genome to which they were aligned (i.e. hard-masked DNA sequences). Thus, the final dataset was composed by 24,830 nuclear SNPs (Table 5).

4.3.1 Identification of individuals of farmed ancestry

A total of 34 individuals of hatchery ancestry were identified with the outlined criteria in the wild populations studied using the whole genomic information. All individuals from VI, CH, AG1, OM, P2 and P3, TE04 and TE14 were considered wild, while, at the other end, most individuals from QB14 were of hatchery ancestry (only four wild individuals). Therefore, this population was removed for further characterization of Iberian populations. One individual of hatchery ancestry from FE, three from LE, five from CE, five from NU04, four from NU14 and two from BL (as reported by Martínez *et al.* 2007), were identified according to the *qW* criterion. However, the populations with hatchery ancestry were in Hardy-Weinberg Equilibrium (HWE; see next section).

All the ~ 25,000 SNPs were ranked by F_{ST} between hatchery (BA14 and S, samples for release and spawners, respectively) *vs* all wild to identify the SNPs with highest diagnostic power to develop a cost-effective and high-resolution tool to elucidate hatchery ancestry. The number of SNPs with $F_{ST} > 0.95$, $F_{ST} > 0.99$ and $F_{ST} = 1$ were 214, 38 and nine, respectively (Table 6). Then, the performance of these different SNP subsets obtained were evaluated as the percentage of correct classification regarding the whole SNP dataset (Fig. 13). In ten populations the three subsets showed the same classification success as the whole SNP dataset. Despite the subset with the highest number of SNPs (214 SNPs panel) showed the best performance, the 38 SNPs panel performed very similarly and only one individual from NU14, excluding CE, was not correctly classified. Furthermore, with only nine diagnostic SNPs the classification success was encouraging. In all evaluations, CE was remarkable because the classification success dropped to ~80% regarding the whole SNP dataset (Fig. 13).

Table 6. Ancestry informative markers (AIMs) selected. W: Wild populations, QB14 population was not employed; H: Hatchery samples, BA14 and S.

F_{ST} *W-H*	Number of SNPs	Number of chromosomes represented
>0.95	214	37
>0.99	38	18
1	9	5

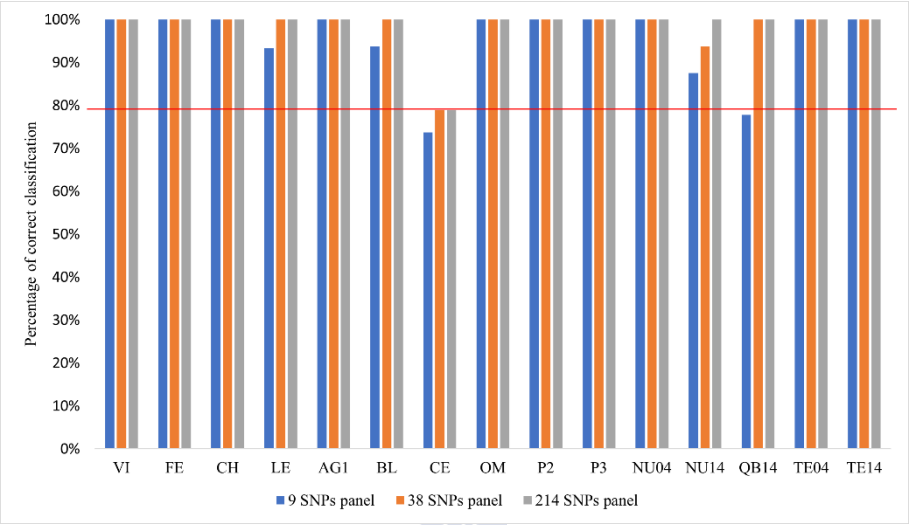

Figure 13. Correct individual hatchery ancestry classification in all populations studied using the whole SNP dataset. The 80% of correct classification is highlighted with a red line.

The lower performance in CE of most SNP subsets has likely to do with the *qW* values of those misclassified individuals, very close to the threshold established, and the same occurred with NU14-01 individual (Table 7). The *qH* obtained among the different SNP subsets showed a highly significant correlation between each SNP subset and the whole SNP dataset, especially when using the panel of 38 and 214 SNPs (r = 0.88 and 0.92, respectively; p-value < 0.001). The incidence of restocking measured as the mean hatchery ancestry (*qH*) was between low and moderate (range: FE 0.01 - NU04 0.11; Table 8) excluding QB14, the most affected population (*qH* = 0.20). The Atlantic Slope

showed a lower incidence of restocking than the Mediterranean one (average *qH*: 0.022 *vs* 0.074, respectively), as previously reported (Almodóvar *et al.* 2006). All populations conformed to Hardy-Weinberg equilibrium (HWE; Fisher's global test *p*-value > 0.05), which suggest full admixture across generations of introgression since the start of restocking programs, although this figure was lower when hatchery ancestry individuals were removed excluding NU04 population (see below HWE).

Table 7. Hatchery ancestry individuals in the populations studied detected with different SNP panels. HAP: Hatchery Allele Proportion; *MD*: Missing data; P: Presence of hatchery alleles in genotype; A: Absence of hatchery alleles in genotype. The *qH* was obtained with different SNP subsets and two populations (K = 2). In red, *qH* > 0.05. The SNPs with F_{ST} = 1 are listed in Table S8 (see Supplementary tables).

	SNPs with F_{ST} = 1 between hatchery and wild populations												
Individuals	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	***HAP***	***qH* (9 SNPS)**	***qH* (38 SNPs)**	***qH* (214 SNPs)**	***qH* (whole panel)**
FE-60	A	A	A	P	A	A	P	P	A	16.7%	0.052	0.240	0.139	0.125
LE-38	*MD*	*MD*	A	*MD*	A	*MD*	*MD*	P	*MD*	33.3%	0.094	0.481	0.391	0.498
LE-41	*MD*	*MD*	A	*MD*	*MD*	*MD*	P	P	P	37.5%	0.244	0.330	0.296	0.446
LE-48	*MD*	A	*MD*	A	*MD*	*MD*	*MD*	P	*MD*	16.7%	0.030	0.069	0.113	0.168
BL-15	P	P	P	P	P	P	P	P	*MD*	81.3%	0.869	0.963	0.989	1.000
BL-24	P	P	P	P	P	*MD*	P	P	*MD*	50.0%	0.495	0.387	0.270	0.279
CE-17	A	A	A	A	A	A	A	A	A	0.0%	0.002	0.001	0.000	0.120
CE-23	*MD*	A	*MD*	A	A	A	A	A	A	0.0%	0.003	0.103	0.132	0.405
CE-27	A	A	A	*MD*	A	A	A	A	A	0.0%	0.002	0.001	0.001	0.144
CE-30	A	P	P	A	A	*MD*	A	A	A	12.5%	0.028	0.003	0.005	0.114
CE-32	A	A	A	A	A	*MD*	A	A	*MD*	0.0%	0.003	0.001	0.000	0.092
NU04-06	P	P	P	P	*MD*	A	A	A	A	25.0%	0.201	0.245	0.265	0.268
NU04-07	P	P	P	P	P	P	P	P	P	50.0%	0.488	0.479	0.491	0.478
NU04-09	A	A	*MD*	P	P	P	P	P	P	37.5%	0.356	0.334	0.309	0.325
NU04-13	P	P	P	P	*MD*	P	P	P	P	50.0%	0.488	0.494	0.479	0.477
NU04-16	P	P	*MD*	A	P	A	A	A	A	18.8%	0.120	0.136	0.164	0.159
NU14-01	A	A	A	A	A	A	P	P	A	11.1%	0.029	0.008	0.069	0.069
NU14-04	P	*MD*	P	A	*MD*	A	P	P	P	50.0%	0.488	0.529	0.423	0.429
NU14-11	A	A	A	P	A	A	P	P	P	38.9%	0.360	0.356	0.270	0.228
NU14-31	A	A	A	A	A	P	A	A	A	33.3%	0.022	0.172	0.255	0.218
QB14-301	P	P	P	P	P	P	P	P	*MD*	68.8%	0.633	0.329	0.252	0.275
QB14-303	A	A	*MD*	P	P	P	A	A	*MD*	28.6%	0.226	0.354	0.442	0.455
QB14-304	P	P	P	A	*MD*	P	P	P	P	50.0%	0.451	0.317	0.262	0.232
QB14-305	P	P	P	P	P	A	*MD*	*MD*	P	42.9%	0.371	0.310	0.334	0.323
QB14-306	P	P	P	A	*MD*	A	P	P	*MD*	35.7%	0.322	0.428	0.393	0.355
QB14-313	P	P	*MD*	A	*MD*	P	A	A	A	21.4%	0.203	0.148	0.204	0.160
QB14-315	P	P	P	P	P	P	P	P	P	50.0%	0.444	0.203	0.176	0.107
QB14-316	A	A	A	A	A	A	P	P	P	16.7%	0.135	0.126	0.118	0.087
QB14-322	A	A	A	P	P	P	P	P	P	33.3%	0.264	0.263	0.230	0.230
QB14-323	P	P	P	A	P	P	A	A	A	44.4%	0.400	0.409	0.362	0.309
QB14-324	A	A	*MD*	P	P	A	A	A	A	25.0%	0.174	0.344	0.426	0.423
QB14-328	A	A	A	P	P	A	P	P	P	44.4%	0.370	0.315	0.292	0.265
QB14-329	P	*MD*	*MD*	A	P	P	P	A	*MD*	41.7%	0.360	0.269	0.252	0.288
QB14-330	P	P	P	P	A	A	P	P	P	38.9%	0.340	0.282	0.223	0.167
HAP	31.7%	26.7%	30.8%	30.6%	32.7%	28.6%	37.1%	36.4%	32.0%					

Table 8. Mean and standard deviation (SD) of hatchery ancestry (*qH)* in introgressed populations from the whole panel STRUCTURE analysis.

Population	**Mean *qH***	**SD *qH***	***Population***	**Mean *qH***	**SD *qH***
FE	0.01	0.04	*NU04*	0.11	0.18
LE	0.07	0.17	*NU14*	0.06	0.12
BL	0.08	0.26	*QB14*	0.20	0.15
CE	0.05	0.10			

Local ancestry inference

STRUCTURE analyses indicated genome-wide admixture in some wild populations affected by restocking, particularly in QB14. Despite four individuals of this population were classified as wild according to the threshold stablished, this does not exclude that a small fraction of the genome might has hatchery ancestry, representing the extreme of the hatchery ancestry distribution in the population. Consequently, the whole population sample was used for the inference of local ancestry (N = 18). A panel composed by 9,640 SNPs with F_{ST} > 0.10 between the two reference genetic pools (i.e. BA14 and TE14) was used. LAMP analyses showed that genomic blocks of potential hatchery ancestry were unevenly spread across the genome (Fig. 14). Some blocks with average hatchery ancestry > 0.50 were identified in three chromosomes. The four individuals classified previously as wild also presented a certain proportion of hatchery ancestry. These results would be expected in a panmictic population heavily affected by hatchery introgression across several generations.

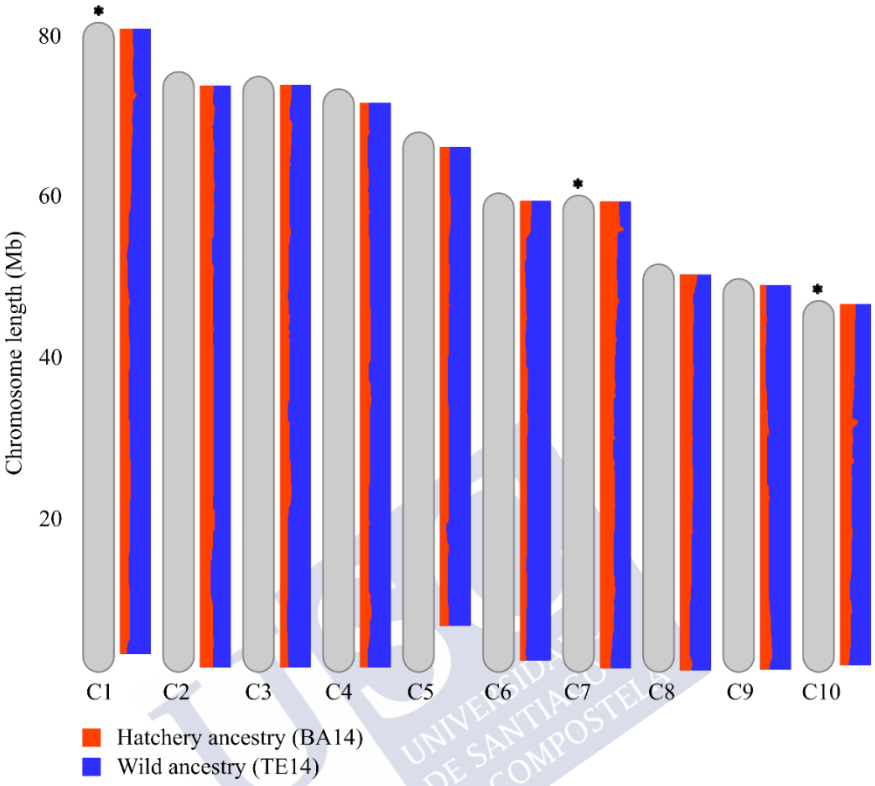

Figure 14. Local average ancestries along *S. trutta* chromosomes (C1-C40), inferred from QB14 population (N = 18). In red and blue, hatchery and wild ancestries, respectively. Both values sum to one. The asterisk indicates chromosomes with regions with average hatchery ancestry > 0.50. A panel with 9,460 SNPs with F_{ST} > 0.10 between reference populations was used.

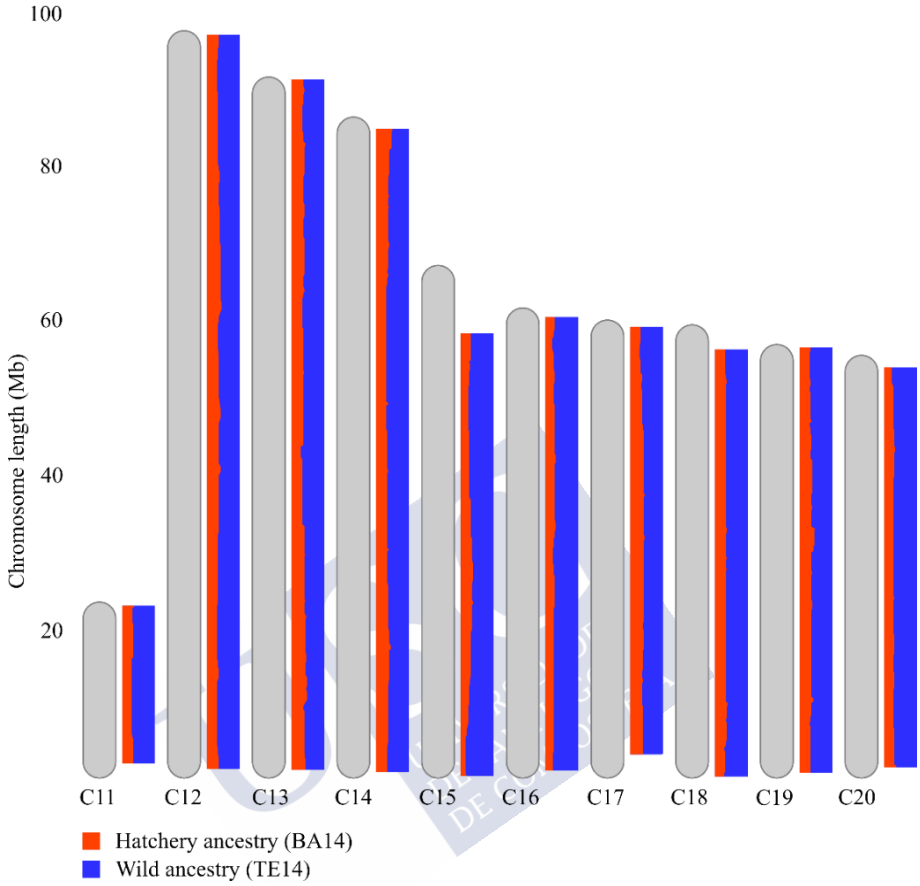

Figure 14. ***(Cont.)***

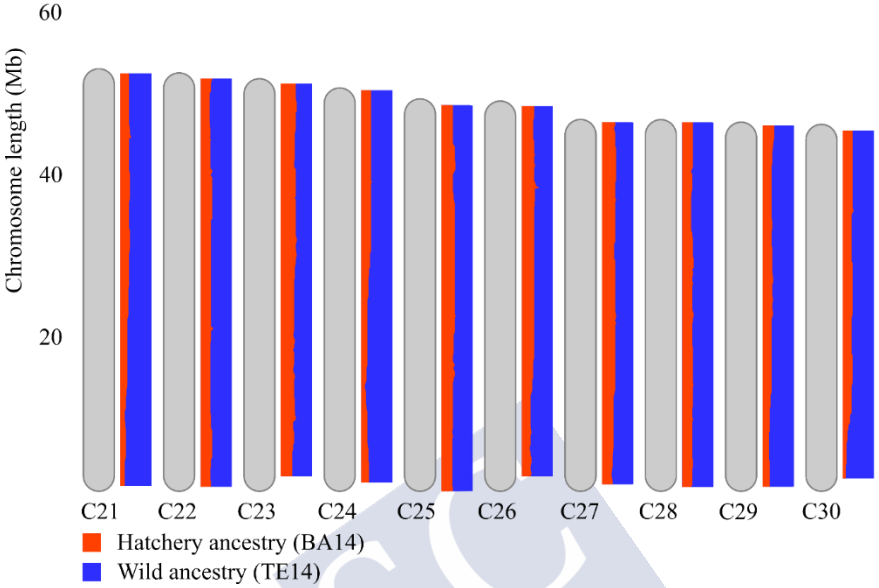

Figure 14. ***(Cont.)***

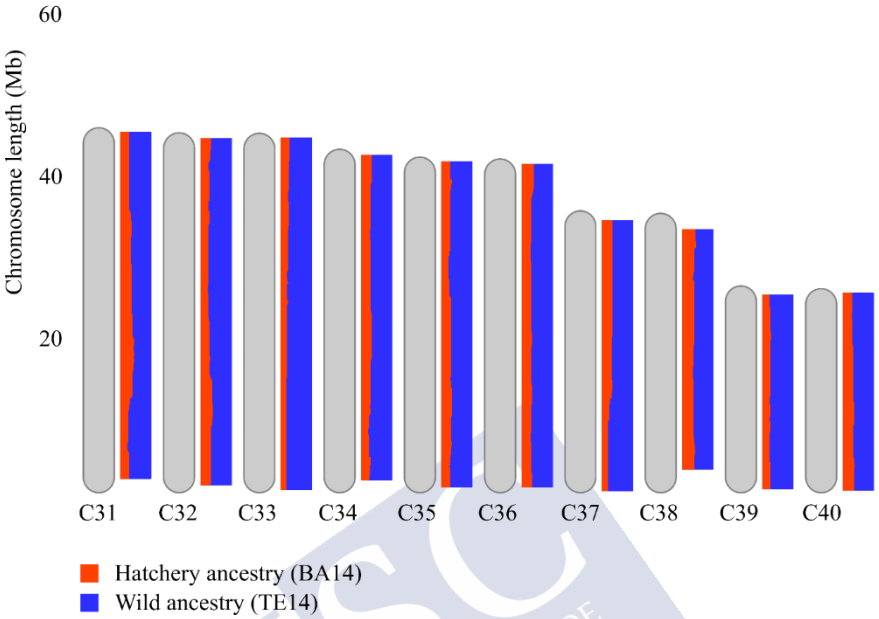

Figure 14. ***(Cont.)***

4.3.2 Genetic diversity

Only wild individuals were used for estimation of genetic diversity and population structure. Genetic diversity estimators were rather heterogeneous across populations and river basins (Table 9). Allelic richness (A_r) ranged from 1.067 (P2) to 1.38 (FE). Expected heterozygosity ranged between Pisuerga river populations (P2: 0.023 and P3: 0.031) and two Miño basin populations (CH: 0.121 and FE: 0.123). Within the Atlantic Slope expected heterozygosity was more than double in Miño than in Duero Basin (average range: 0.106 *vs* 0.045), while the Ter Basin in the Mediterranean Slope showed the lowest value (0.040), being slightly lower than Duero Basin. Between temporal replicates the highest difference was found in Núria River (H_e 0.035 *vs* 0.044 for NU04 and NU14, respectively). Observed and expected heterozygosities were very similar in each population as reflected by the intrapopulation fixation index (F_{IS}), which was low but negative for almost all populations. These values were significant in most populations using the 95% confidence interval approach ($F_{IS} \neq 0$; value not included in the confidence interval), despite all of them did not globally deviated from random mating according to exact tests (see below Hardy-Weinberg equilibrium). Both hatchery stock samples showed high genetic diversity figures, very close to the upper range detected in Miño Basin. A notable amount of private alleles were detected, in all cases at very low frequencies, reflecting the structure of brown trout populations.

Table 9. Genetic diversity in brown trout populations from Iberian Peninsula and a hatchery stock. A_r, mean allelic richness; H_o, mean observed heterozygosity; H_e, mean expected heterozygosity; F_{IS}, global intrapopulation fixation index; BC 95% CI, bias corrected 95% confidence intervals of F_{IS} values; PA, number of private alleles. No hatchery ancestry individuals were used, with the exception of hatchery individuals from Bagà fish farm (i.e. BA14 and S). A total of 245 individuals were used. The values were rounded to three decimal numbers.

Population codes	A_r	H_o	H_e	F_{IS}	lower and upper BC 95% CI	PA
VI	1.239	0.088	0.080	*-0.100*	-0.155 -0.065	340
FE	1.380	0.134	0.123	*-0.085*	-0.148 -0.047	119
CH	1.358	0.129	0.121	*-0.065*	-0.124 -0.031	170
LE	1.321	0.111	0.101	*-0.102*	-0.190 -0.050	68
AG1	1.135	0.045	0.045	*0.001*	-0.086 0.104	457
BL	1.276	0.097	0.091	*-0.069*	-0.114 -0.043	471
CE	1.203	0.047	0.043	*-0.093*	-0.158 -0.050	179
OM	1.112	0.038	0.035	*-0.075*	-0.102 -0.053	316
P2	1.067	0.024	0.023	*-0.050*	-0.085 -0.023	86
P3	1.090	0.035	0.031	*-0.143*	-0.232 -0.089	156
BA14	1.275	0.09	0.085	*-0.059*	-0.089 -0.038	197
S	1.301	0.098	0.090	*-0.084*	-0.144 -0.040	312
NU04	1.105	0.037	0.035	*-0.067*	-0.185 0.022	35
NU14	1.152	0.047	0.044	*-0.067*	-0.134 -0.021	37
TE4	1.120	0.043	0.041	*-0.041*	-0.091 -0.004	10
TE14	1.114	0.041	0.039	*-0.061*	-0.107 -0.018	17

No deviations from Hardy-Weinberg equilibrium (HWE) were detected at population level (global Fisher's test; Table 10). The proportion of loci that showed HWE deviation at *p*-value < 0.05 ranged from 0.9% in NU04 to 10.7% in AG1 with an average of 3.0%, below to expected 5% expected by chance. The proportion of deviations increased when the individuals with hatchery ancestry were included in the analysis, as expected due to a Wahlund effect.

Table 10. Deviation from Hardy-Weinberg (HW) expectations for each population over all loci using Fisher's exact tests. In grey when individuals with hatchery ancestry were included in analysis.

Population codes	No. individuals	N SNPs per analysis	HW (global *p*-value)	HW (% *p*-values <0.05)
VI	15	5,260	1.00	2.1
FE	12	8,006	1.00	1.5
FE^H	13	8,314	1.00	1.8
CH	13	8,033	1.00	1.6
LE	12	6,166	1.00	1.1
LE^H	15	7,193	1.00	1.9
AG1	16	3,145	1.00	10.7
BL	14	6,192	1.00	1.8
BL^H	15	6,693	1.00	4.3
CE	14	3,448	1.00	1.1
CE^H	19	6,510	1.00	2.7
OM	20	2,479	1.00	3.0
P2	18	1,609	1.00	3.5
P3	17	2,119	1.00	5.5
BA14	19	6,923	1.00	3.0
S	20	7,686	1.00	1.9
NU04	11	2,247	1.00	1.3
$NU04^H$	16	7,679	1.00	0.9
NU14	12	2,751	1.00	2.3
$NU14^H$	16	6,374	1.00	10.7
$QB14^H$	18	9,690	1.00	1.8
TE04	14	2,790	1.00	2.0
TE14	18	2,754	1.00	2.9

4.3.3 Genetic differentiation and structure

All SNPs were included in these analyses. All pairwise F_{ST} comparisons were significant (Table 11), no confidence intervals included zero. Using wild and hatchery populations, the lowest pairwise F_{ST} was found between BA14 and S (hatchery) and between temporal replicates from Ter (F_{ST} = 0.015 and F_{ST} = 0.017, for BA14-S and TE04-TE14 respectively). Higher differentiation values were obtained in Núria temporal replicates. Between Núria replicates, F_{ST} value was far higher when exclusively wild samples were used (F_{ST} from 0.030 to 0.102 when using all and wild individuals, respectively). High pairwise differentiation between populations belonging to the same river basin was observed in Duero Basin (F_{ST} = 0.607 between AG1-P2), including populations that belonged to the same tributary (i.e. Pisuerga River, F_{ST} = 0.294 between P2-P3). Putative hybrid populations from Miño Basin showed lower F_{ST} values with DU lineage population (i.e. LE) than with the AT representative population (VI). The global F_{ST} was 0.680 (*p*-value = 0.000).

Pairwise basin comparisons between Mediterranean (Ter) and the Atlantic Slope (Miño and Duero) were higher (F_{ST} values > 0.660; Table 12) than within Atlantic Slope (F_{ST} = 0.300). Global F_{ST} was higher in Duero Basin than in Miño Basin (0.460 *vs* 0.192, respectively).

Table 11. Pairwise F_{ST} values between populations (below diagonal) obtained with N = 245 individuals from wild and hatchery populations and the whole SNP panel with 24,830 SNPs.

	Miño Basin					Duero Basin					Hatchery		Mediterranean Basin			
	VI	FE	CH	LE	AG1	BL	CE	OM	P2	P3	BA14	S	NU04	NU14	TE04	TE14
VI																
FE	0.265															
CH	0.267	0.050														
LE	0.326	0.071	0.106													
AG1	0.584	0.419	0.445	0.425												
BL	0.403	0.218	0.243	0.222	0.401											
CE	0.575	0.402	0.429	0.407	0.536	0.431										
OM	0.627	0.466	0.494	0.458	0.538	0.471	0.372									
P2	0.671	0.515	0.539	0.522	0.607	0.523	0.428	0.390								
P3	0.636	0.474	0.501	0.470	0.556	0.481	0.340	0.288	0.294							
BA14	0.567	0.473	0.475	0.552	0.703	0.584	0.695	0.739	0.762	0.743						
S	0.557	0.466	0.468	0.543	0.692	0.576	0.684	0.728	0.750	0.731	0.015					
NU04	0.784	0.708	0.709	0.761	0.856	0.764	0.860	0.876	0.903	0.886	0.735	0.721				
NU14	0.768	0.693	0.694	0.744	0.840	0.750	0.842	0.862	0.887	0.871	0.719	0.706	0.102			
TE04	0.780	0.711	0.711	0.759	0.846	0.762	0.848	0.866	0.889	0.874	0.732	0.719	0.194	0.156		
TE14	0.793	0.732	0.732	0.777	0.852	0.778	0.854	0.869	0.890	0.876	0.747	0.734	0.223	0.188	0.017	

Table 12. Pairwise F_{ST} values between river drainages (below diagonal). F_{ST} values were obtained with N = 176 individuals from wild contemporary populations and the 20,293 SNPs panel. Atlantic Slope: Miño Basin populations: VI, FE, CH, LE; Duero Basin: AG1, BL, CE, OM, P2, P3; Mediterranean Slope: Ter Basin: NU04, TE04. The global F_{ST} values for each basin are shown on the diagonal.

	MIÑO	DUERO	MED
MIÑO	0.192		
DUERO	0.300	0.460	
MED	0.661	0.771	0.151

AMOVA results revealed significant genetic structuring among river basins (Table 13) belonging to Atlantic *vs* Mediterranean slopes or between different river basins. However, only marginal significant genetic structuring was found in Duero Basin according to two of the criteria established and no intergroup (F_{CT}) significant values were detected in Miño Basin in the two scenarios tested. In the Mediterranean Slope, the geographical variance (locations) within Ter Basin was higher than temporal variance (temporal samples). In all hypotheses tested the highest variance was found within populations.

Table 13. Analyses of AMOVA with Arlequin 3.5.2.2. In light green, grouping hypotheses with populations coming from different river basins. Atlantic Slope with populations from Miño and Duero basins and Mediterranean Slope with two samples (NU04 and TE04) from Ter Basin. In dark yellow, population structure between localities and temporal replicates in Mediterranean Slope. In light yellow, hypotheses tested according to STRUCTURE results and previous reports using populations from Duero Basin (Bouza *et al*. 2001; Martínez *et al*. 2007). Atlantic mtDNA lineage (*AT*): AG1 and BL populations. Putative hybrid zone (*PHZ*: CE and OM populations). Duero mtDNA lineage (*DU*): P2 and P3 populations. In light blue, hypothesis tested using populations from Miño Basin according to STRUCTURE results and previous references (Bouza *et al*. 2008 and Vilas *et al*. 2010).

Hypotheses	df	Variance	% Variation
1. Two groups (Atlantic *vs* Mediterranean Slope)			
Among groups	1	1994.74	67.35
Among populations within groups	10	420.40	14.19
Within populations	340	546.79	18.46
F statistics: F_{CT} = 0.67**, F_{SC} = 0.43***, F_{ST} = 0.82***			
2. Three groups (Miño, Duero and Ter basins)			
Among groups	2	1064.83	56.26
Among populations within groups	9	281.162	14.85
Within populations	340	546.79	28.89
F statistics: F_{CT} = 0.56***, F_{SC} = 0.34***, F_{ST} = 0.71***			
3. Two groups (Miño *vs* Duero basins)			
Among groups	1	298.05	25.69
Among populations within groups	8	298.91	25.77
Within populations	292	563.11	48.54
F statistics: F_{CT} = 0.26**, F_{SC} = 0.35***, F_{ST} = 0.51***			

+*p*-value < 0.1, ****p*-value < 0.05, *****p*-value < 0.01.

Table 13. ***(Cont.)***

Hypotheses	df	Variance	% Variation
4. Two groups within MED Slope (NU04+NU14 *vs* TE04+TE14)			
Among groups (rivers)	1	80.46	14.65
Among temporal replicates (2004-14)	2	24.89	4.53
Within populations	106	443.99	80.82
F statistics: F_{CT} = 0.15, F_{SC} = 0.05***, F_{ST} = 0.19***			
5a. Two groups within Duero Basin (AT *vs* PHZ+DU)			
Among groups	1	248.54	27.52
Among populations within groups	4	233.32	25.84
Within populations	192	421.15	46.64
F statistics: F_{CT} = 0.28^{+}, F_{SC} = 0.36***, F_{ST} = 0.53***			
5b. Three groups within Duero Basin (AT, PHZ, DU)			
Among groups	2	142.72	17.64
Among populations within groups	3	245.40	30.32
Within populations	192	421.15	52.04
F statistics: F_{CT} = 0.176^{+}, F_{SC} = 0.37***, F_{ST} = 0.48***			
5c. Four groups within Duero Basin (*AT*, PHZ, P2, P3)			
Among groups	3	49.75	6.32
Among populations within groups	2	316.13	40.17
Within populations	192	421.152	53.51
F statistics: F_{CT} = 0.06, F_{SC} = 0.43***, F_{ST} = 0.46***			

+*p*-value < 0.1, ****p*-value < 0.05, *****p*-value < 0.01.

Table 13. ***(Cont.)***

Hypotheses	df	Variance	% Variation
6a. Two groups within Miño Basin (VI, FE+CH+LE)			
Among groups	1	225.89	20.68
Among populations within groups	2	44.35	4.06
Within populations	100	821.93	75.26
F statistics: F_{CT} = 0.21, F_{SC} = 0.051***, F_{ST} = 0.25 ***			
6b. Three groups within Miño Basin (VI, FE+CH, LE)			
Among groups	2	136.54	13.50
Among populations within groups	1	52.88	5.23
Within populations	100	821.93	81.27
F statistics: F_{CT} = 0.14, F_{SC} = 0.06***, F_{ST} = 0.19***			

+*p*-value < 0.1, ****p*-value < 0.05, *****p*-value < 0.01.

Atlantic and Mediterranean slopes

When populations from both slopes were considered, the most likely number of populations units varied among K estimators: ΔK, MedMean K and Mean LnP(K) rendered two, six and ten as the most probable K values, respectively. With the lowest K, two clusters were identified including populations belonging to Atlantic and Mediterranean slopes, respectively (Fig. 15). The Mediterranean cluster was consistent across all K values, while the Atlantic cluster was progressively subdivided as the number of Ks increased. However, the difference between K = 6 and K = 10 was related only to BL, which in the former case appeared to be constituted by two genomic components, one BL specific but the other related to the inner Miño Basin. Excluding this population, the Duero and Miño basins appeared roughly divided

in two clusters related to outlet (AG1 and VI) and inner populations. Interestingly, one population in both river basins appeared to be admixed with opposite frequencies of the same components (CE and LE). Finally, an individual with a genome constitution pertaining to the inner cluster in Duero basin was clearly visible with this analysis in AG1 population (AG1-22), a fact confirmed by applying GENECLASS2 (p-value < 0.01).

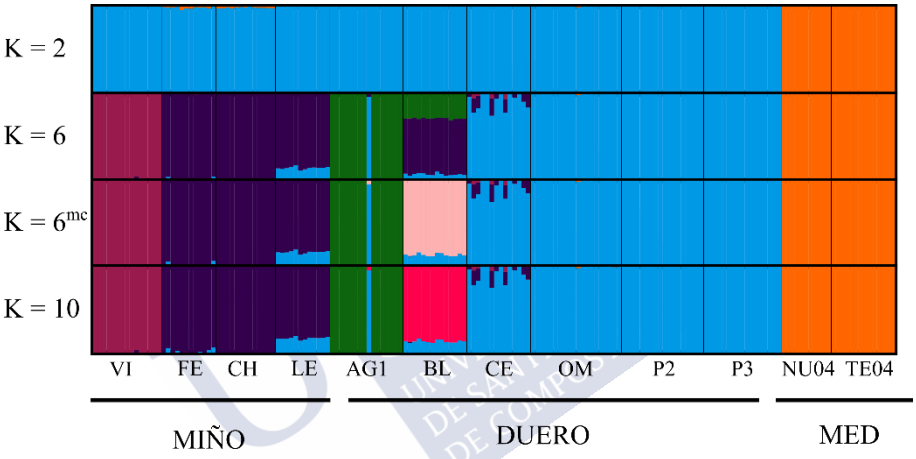

Figure 15. CLUMPAK plot of STRUCTURE assignment with different Ks, (K = 2,6 and 10). N = 176 wild individuals from twelve populations and 20,293 SNPs were used. mc: minor cluster. Each individual is represented as a vertical bar partitioned into segments according to the proportion of the genome belonging to each of the clusters identified (K) by STRUCTURE.

For DAPC analyses, the lowest BIC value (1226.60) corresponded to K = 9, although adjacent Ks showed rather similar values. The individual cluster assignment with K = 12 was the same as that defined *a priori* according to sampling sites, excluding three individuals (VI-19 and CH-43 assigned to LE population and AG1-22 assigned to the inner Duero cluster). After the cross-validation method was applied, 23 PCA axes and four discriminant functions were retained (76.1% of the variance). To retain more than 90% of the variance, 85 PCA axes and

five discriminant functions would be necessary comprehending 90.1% of the variance. In all scatterplots, populations from Mediterranean slope constituted a clearly differentiated cluster, while the Atlantic slope showed a much lesser consistent pattern regarding geography or previous reports with other genetic markers (Fig. 16).

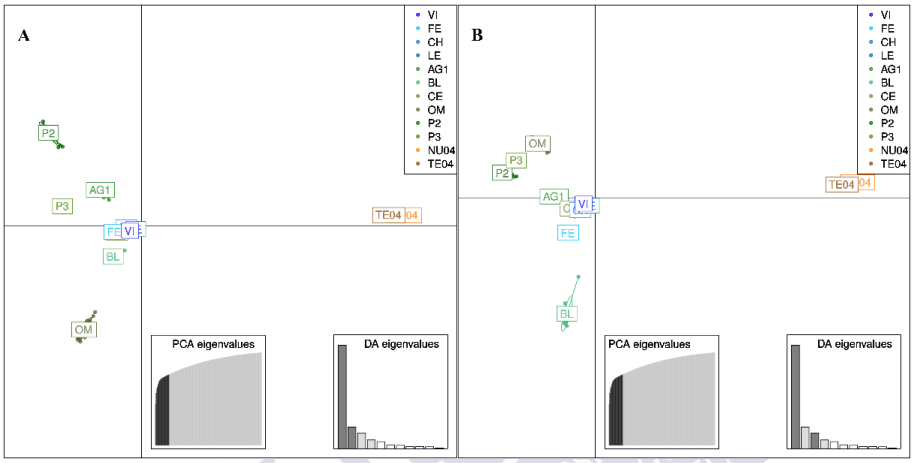

Figure 16. Scatterplots of individuals on the three principal DA eigenvalues of DAPC. Populations had three different basic colours: blue (Miño Basin), green (Duero Basin) and brown (Mediterranean populations). The graph represents the individuals as dots and the groups as inertia ellipses. PCA and DA eigenvalues are displayed inset. A: Scatterplot with first and second DA eigenvalues. B: Scatterplot with first and third DA eigenvalues.

Atlantic Slope (Miño and Duero basins)

When populations from Mediterranean Slope were excluded the picture in the Atlantic Slope barely changed. The most likely number of units varied among K estimators: ΔK, MedMean K and Mean LnP(K) were two, six and ten, respectively. With the lowest K, the clusters obtained roughly separated Miño and Duero basins (Fig. 17), excluding BL whose composition was closer to Miño than to Duero Basin. These two clusters (orange and blue) were consistently maintained across all K values with some refinement related to the singularity of the outlet populations (VI and AG1), also observed with

the whole dataset (Fig. 15), but with a differentiation in the inner part of Duero Basin associated to OM and to a minor extent to CE with K = 10.

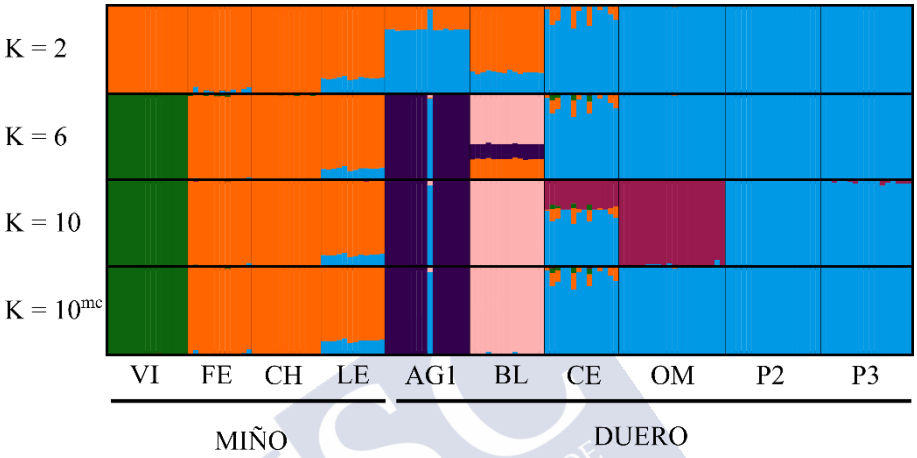

Figure 17. CLUMPAK plot of STRUCTURE assignment with different Ks (K = 2, 6 and 10). N = 151 wild individuals from ten populations from Atlantic Slope and 15,158 SNPs were used. mc: minor cluster. Each individual is represented as a vertical bar partitioned into segments according to the proportion of the genome belonging to each of the clusters identified (K) by STRUCTURE.

For DAPC analyses, the lowest BIC value (1045.69) corresponded to K = 9, although adjacent Ks showed quite similar values. The individual cluster assignment with K = 10 was the same as that defined *a priori* according to sampling sites with only two exceptions (CH-43 assigned to LE population and AG1-22 to the inner Duero Basin). After the cross-validation method, 13 PCA axes and five discriminant functions were retained (57.8% of the variance). To retain more than 90% of the variance, 90 PCA axes and four discriminant functions would be necessary comprehending 91.1% of the variance. In all scatterplots, the most consistent result is the notable differentiation of BL from all other populations (Fig. 18).

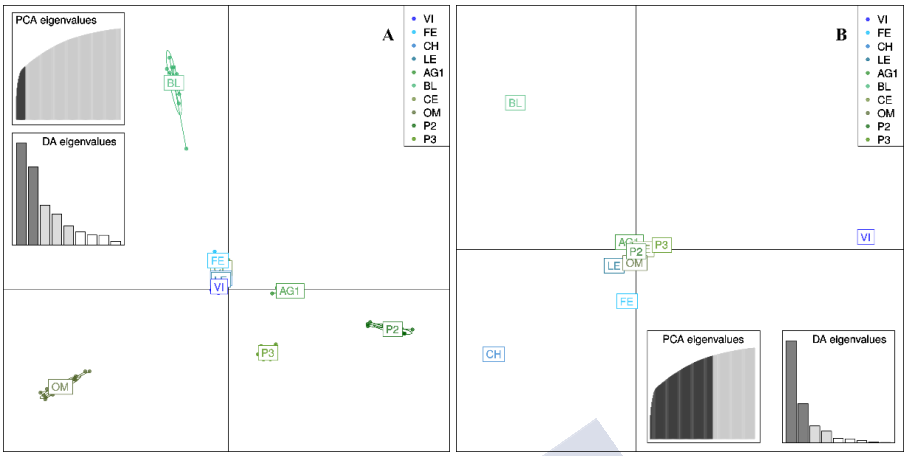

Figure 18. Scatterplots of individuals on the two principal DA eigenvalues of DAPC. Populations are coloured according to their basin, blue shades for Miño and green shades for Duero. The graph represents the individuals as dots and the groups as inertia ellipses. PCA and DA eigenvalues are displayed inset. A: Scatterplot with 57.8% of the variance. B: Scatterplot with 91.1% of the variance.

Duero Basin

The most likely number of populations units was similar among K estimators: MedMean K and ΔK, Mean LnP(K), five and six, respectively. By increasing the values of K more refined information about populations structure is observed (Fig. 19). With K = 2, two clusters were delineated corresponding to the lower-course (AG1-BL) and the inner part of Duero Basin. Higher Ks determined the breaking down of some populations from the initial clusters (BL, CE, and OM), although CE always maintained a mixed constitution, and P3 also showed mixed composition at higher Ks. Two reference populations consistently maintained across all Ks, AG1 in the outlet and P2 in the inner part.

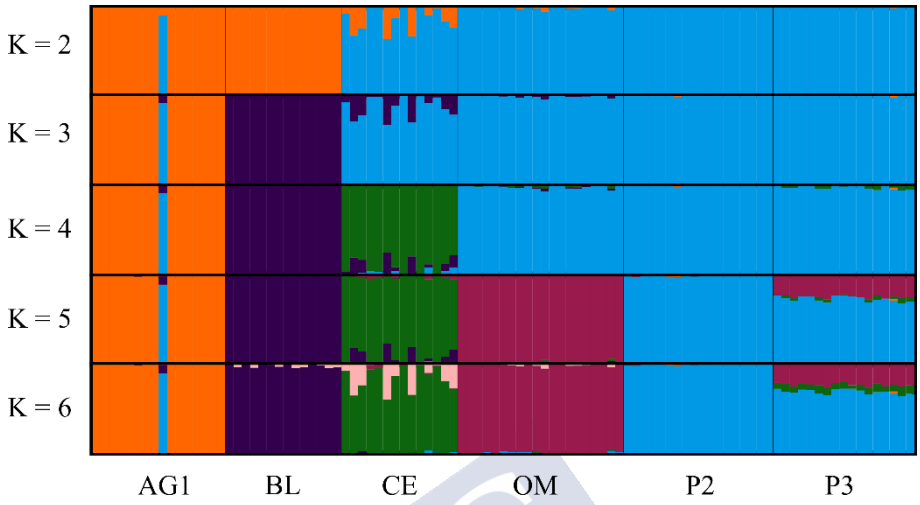

Figure 19. CLUMPAK plot of STRUCTURE assignment from K = 2 to 6. N = 99 wild individuals from six populations from Duero Basin and 10,777 SNPs were used. Each individual is represented as a vertical bar partitioned into segments according to the proportion of the genome belonging to each of the clusters identified (K) by STRUCTURE.

For DAPC analyses, the lowest BIC value (640.58) corresponded to K = 6. The individual cluster assignment with K = 6 was the same as that defined *a priori* according to sampling sites with only one exception (AG1-22 assigned to the inner Duero Basin). After the cross-validation method was applied, eight PCA axes and four discriminant functions were retained (60.6% of the variance). To retain more than 90% of the variance, 60 PCA axes and three discriminant functions would be necessary comprehending 91.0% of the total variance. In all scatterplots, BL was consistently separated from the remaining populations that did not show any remarkable pattern (Fig. 20).

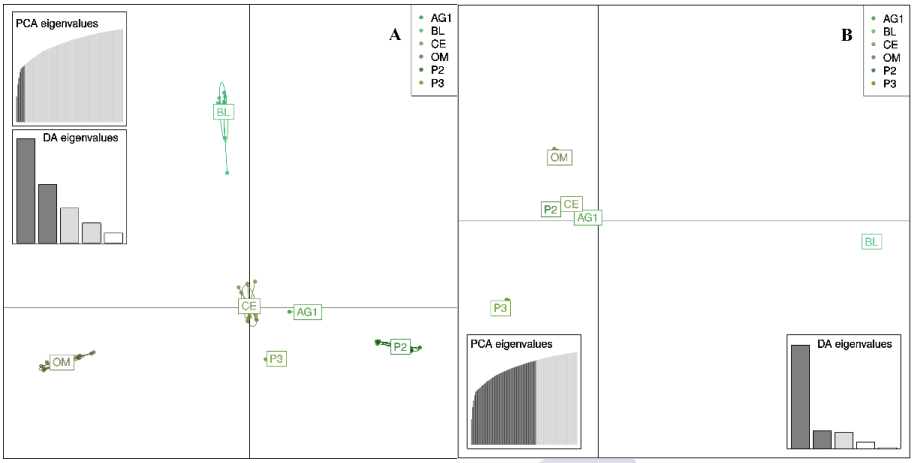

Figure 20. Scatterplots of individuals on the two principal DA eigenvalues of DAPC. All populations belong to Duero Basin. The graph represents the individuals as dots and the groups as inertia ellipses. PCAs and DAs eigenvalues are displayed inset. A: Scatterplot with 60.6% of the variance. B: Scatterplot with 91.0% of the variance.

Wild hybrid zones (Duero and Miño basins)

Within the Duero Basin, the most likely number of populations units varied among K estimators: ΔK, MedMean K, and Mean LnP(K), two, five, and six, respectively. With K = 2, BL appeared as a distinct mixed population, not so CE (Fig. 21). Once more, the results obtained do not show OM as a hybrid population. With the highest K each population was represented as a separate cluster.

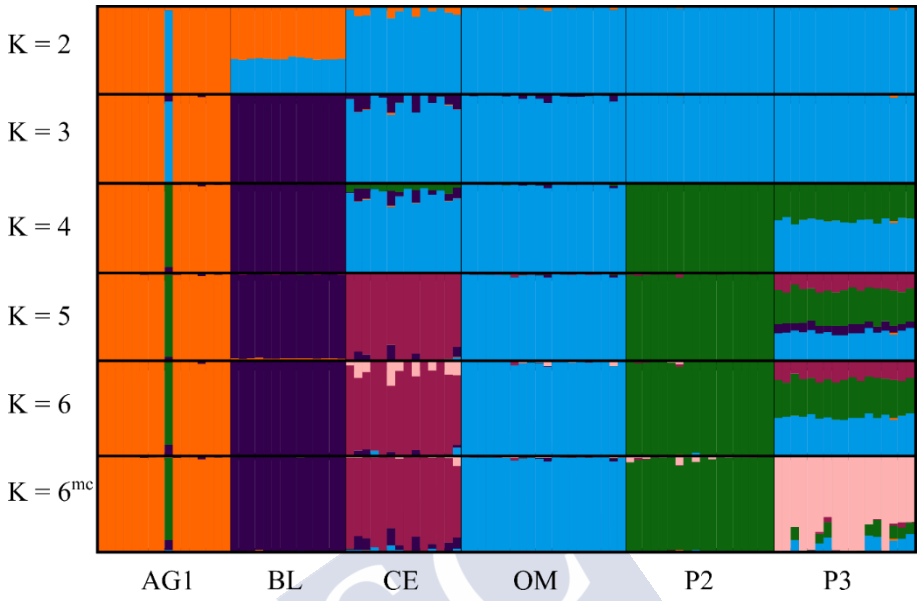

Figure 21. CLUMPAK plot of STRUCTURE assignment from K = 2 to 6. N = 99 wild individuals from six populations from Duero Basin and 1,778 AIMs were used. mc: minor cluster. Each individual is represented as a vertical bar partitioned into segments according to the proportion of the genome belonging to each of the clusters identified (K) by STRUCTURE.

Within the Miño Basin, the most likely number of populations units was similar among K estimators: ΔK with two and MedMean K, and Mean LnP(K) with three. With K = 2, two clusters were delineated corresponding to VI in the outlet and the inner Miño Basin populations, with some degree of hybridization observed in FE and CH ($q_{VI} \sim 0.10$; Fig. 22). With K = 3, FE and CH appeared to be hybrid populations but one of the components would not belong to any of the reference populations used.

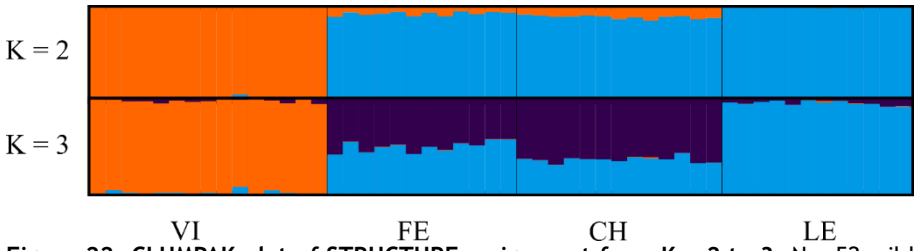

Figure 22. CLUMPAK plot of STRUCTURE assignment from K = 2 to 3. N = 52 wild individuals from four populations from Miño Basin and 1,429 AIMs were used. Each individual is represented as a vertical bar partitioned into segments according to the proportion of the genome belonging to each of the clusters identified (K) by STRUCTURE.

4.3.4 Effective population size

The estimates of effective population size (N_e) using NeEstimator software yielded quite often finite values (Tables 14 and 15). A plateau in N_e estimates was obtained with some populations (e.g P2 and P3). Nevertheless, the upper boundaries of the 95% confidence intervals were infinite with many populations and the different MAF thresholds applied, indicating that the estimates may not be very robust in these cases. N_e estimates of the same order of magnitude were obtained between LDN_e and temporal methods when temporal replicates were available, showing that Ter population would have a higher effective population size than Núria population (Tables 14 and 16), the last one with higher hatchery ancestry proportion. N_e estimates using pairs of SNPs placed in distinct chromosomes showed higher values (Table 15) than obtained considering all the markers available, which makes sense since LD is reduced when using markers on different chromosomes. In any case, the values obtained between both methodologies were similar.

Table 14. Effective population size calculated with the Linkage Disequilibrium method (Waples 2006) using all SNP pairs and considering different MAF thresholds. In parentheses, 95% CI. Inf.: Infinite.

	≥0.10	*≥0.15*	*≥0.20*	*≥0.25*	*≥0.30*	*≥0.35*	*≥0.40*
VI	82.1 (4.7-Inf.)	83.0 (4.1-Inf.)	96.2 (3.8-Inf.)	116.1 (4.0-Inf.)	197.7 (4.5-Inf.)	559.6 (4.6-Inf.)	Inf. (5.5-Inf)
FE	35.1 (9.6-Inf.)	37.7 (9.8-Inf.)	40.2 (9.9-Inf.)	44.0 (10.0-Inf.)	48.7 (9.8-Inf.)	57.9 (10.3-Inf.)	73.8 (10.7-Inf.)
CH	Inf. (6.7-Inf.)	7032.0 (6.0-Inf.)	5607.3 (5.4-Inf.)	160297.9 (5.3-Inf.)	Inf. (5.2-Inf.)	Inf. (4.9-Inf.)	Inf. (5.5-Inf.)
LE	Inf. (1.9-Inf.)	Inf. (1.9-Inf.)	Inf. (1.8-Inf.)	Inf. (1.7-Inf.)	Inf. (1.7-Inf.)	Inf. (1.7-Inf.)	Inf. (1.9-Inf.)
AG1	19.0 (6.7-Inf.)	21.2 (7.6-Inf.)	21.7 (7.7-Inf.)	22.1 (7.7-Inf.)	23.7 (8.1-Inf.)	24.0 (8.1-Inf.)	28.1 (8.4-Inf.)
BL	496.2 (24.0-Inf.)	562.2 (22.2-Inf.)	484.4 (21.9-Inf.)	443.4 (21.6-Inf.)	582.9 (21.2-Inf.)	1301.1 (22.7-Inf.)	Inf. (21.3-Inf.)
CE	3.9 (1.8-20.1)	9.1 (2.6-105.8)	18.8 (4.9-Inf.)	33.5 (6.7-Inf.)	117.4 (9.8-Inf.)	Inf. (12.1-Inf.)	Inf. (15.0-Inf.)
OM	58.1 (23.1-Inf.)	56.0 (22.0-Inf.)	57.5 (22.4-Inf.)	57.2 (22.3-Inf.)	58.7 (22.6-Inf.)	60.0 (23.0-Inf.)	62.9 (23.4-Inf.)
P2	33.0 (18.0-105.9)	33.5 (17.8-119.1)	32.5 (17.8-101.3)	34.9 (18.8-120.3)	37.9 (19.5-167.2)	41.1 (18.8-743.2)	46.6 (19.3-Inf.)
P3	10.4 (6.4-17.8)	11.3 (7.3-18.6)	11.9 (7.7-19.8)	11.7 (7.3-20.2)	11.9 (7.3-21.2)	11.6 (6.9-21.9)	10.1 (6.0-18.2)
NU04	4.2 (2.1-13.2)	3.3 (2.0-11.9)	3.1 (1.9-10.5)	3.0 (1.8-9.8)	2.8 (1.7-10.4)	2.9 (1.7-10.8)	4.6 (2.1-15.0)
NU14	27.0 (8.3-Inf.)	35.0 (10.1-Inf.)	36.9 (10.3-Inf.)	35.7 (10.1-Inf.)	34.2 (9.2-Inf.)	34.7 (9.1-Inf.)	43.4 (10.5-Inf.)
TE04	59.1 (26.2-Inf.)	55.1 (23.1-Inf.)	52.9 (23.6-Inf.)	51.3 (23.4-Inf.)	45.0 (21.6-507.8)	46.1 (22.7-383.9)	39.1 (19.7-226.3)
TE14	24.1 (14.0-56.2)	22.9 (13.5-51.5)	21.4 (12.6-46.8)	22.0 (13.2-46.6)	20.5 (12.2-43.7)	19.4 (11.5-41.1)	18.0 (10.1-42.3)

Table 15. Effective population size calculated with the Linkage Disequilibrium method (Waples 2006) using pairwise comparisons between SNPs located in different chromosomes and considering different MAF thresholds. In parentheses, 95% CI. Inf.: Infinite.

	≥0.10	*≥0.15*	*≥0.20*	*≥0.25*	*≥0.30*	*≥0.35*	*≥0.40*
VI	99.7 (4.8-Inf.)	101.4 (4.3-Inf.)	121.6 (4.0-Inf.)	158.2 (4.2-Inf.)	380.9 (4.6-Inf.)	Inf. (4.8-Inf.)	Inf. (5.6-Inf.)
FE	37.0 (9.8-Inf.)	39.9 (10.0-Inf.)	42.9 (10.1-Inf.)	47.6 (10.2-Inf.)	53.7 (10.0-Inf.)	66.3 (10.5-Inf.)	88.9 (11.0-Inf.)
CH	Inf. (6.8-Inf.)	Inf. (6.1-Inf.)	Inf. (5.5-Inf.)	Inf. (5.4-Inf.)	Inf. (5.2-Inf.)	Inf. (4.9-Inf.)	Inf. (5.5-Inf.)
LE	Inf. (1.9-Inf.)	Inf. (1.9-Inf.)	Inf. (1.8-Inf.)	Inf. (1.7-Inf.)	Inf. (1.7-Inf.)	Inf. (1.7-Inf.)	Inf. (1.9-Inf.)
AG1	20.3 (6.9-Inf.)	23.1 (7.9-Inf.)	23.8 (8.1-Inf.)	24.4 (8.0-Inf.)	26.8 (8.5-Inf.)	28.0 (8.7-Inf.)	34.5 (9.1-Inf.)
BL	4592.4 (25.3-Inf.)	Inf. (23.3-Inf.)	3874.7 (22.9-Inf.)	3665.2 (22.7-Inf.)	Inf. (22.3-Inf.)	Inf. (24.0-Inf.)	Inf. (22.5-Inf.)
CE	6.6 (2.1- 57.6)	12.6 (2.9-Inf.)	27.2 (6.0-Inf.)	59.4 (7.8-Inf.)	Inf. (11.5-Inf.)	Inf. (14.4-Inf.)	Inf. (18.6-Inf.)
OM	64.5 (24.3-Inf.)	63.1 (23.2-Inf.)	66.3 (23.9-Inf.)	66.6 (23.8-Inf.)	68.9 (24.2-Inf.)	72.3 (24.8-Inf.)	77.8 (25.5-Inf.)
P2	38.3 (19.5- 182.9)	39.6 (19.5- 243.6)	39.3 (19.8- 205.9)	43.3 (21.1- 329.5)	48.4 (22.3- 1522.4)	54.4 (21.6-Inf.)	68.4 (23.1-Inf.)
P3	11.2 (7.0-19.1)	12.3 (8.0-20.4)	13.0 (8.4-21.9)	12.7 (8.0-22.4)	12.8 (7.9-23.5)	12.6 (7.3-24.6)	10.9 (6.4- 20.3)
NU04	4.4 (2.1-13.9)	3.3 (2.0-12.6)	3.2 (2.0- 11.1)	3.1 (1.9-10.5)	2.9 (1.7-11.2)	3.0 (1.7-11.6)	5.2 (2.2-16.7)
NU14	29.9 (8.7-Inf.)	39.7 (10.6-Inf.)	42.9 (11.0-Inf.)	41.9 (10.8-Inf.)	41.3 (9.8-Inf.)	43.4 (9.8-Inf.)	56.9 (11.0-Inf.)
TE04	65.2 (27.5-Inf.)	61.3 (24.3-Inf.)	59.7 (25.0-Inf.)	58.4 (24.9-Inf.)	50.4 (23.0-Inf.)	52.3 (24.1-13273.5)	43.1 (20.6- 490.4)
TE14	26.0 (14.8- 65.3)	25.0 (14.4- 60.4)	23.5 (13.6- 55.5)	24.5 (14.3- 56.7)	22.9 (13.3- 53.1)	21.8 (12.6- 50.3)	20.6 (11.2- 54.9)

Table 16. Effective population size calculated with the temporal method. In parentheses, 95% CI (loci jackknife method). With ten years between samples the most plausible number of generations in brown trout would be three (in bold).

Núria (2004-2014)	*Pollak*	*Nei/Tajima*	*Jorde/Ryman*
1 generation	2.5 (2.3-2.6)	3.1 (2.9-3.3)	2.3 (2.1-2.4)
2 generations	4.9 (4.6-5.2)	6.2 (5.8-6.5)	4.5 (4.3-4.8)
3 generations	**7.4 (6.9-7.8)**	**9.2 (8.7-9.8)**	**6.8 (6.4-7.2)**
Ter (2004-2014)			
1 generation	15.1 (13.1-17.6)	18.8 (16.1-22.4)	14.8 (12.7-17.7)
2 generations	30.3 (26.2-35.3)	37.7 (32.2-44.8)	29.6 (25.5-35.3)
3 generations	**45.4 (39.3- 52.9)**	**56.5 (48.3- 67.2)**	**44.4 (38.2-53.0)**

4.3.5 Adaptative variation

In all cases studied, a higher number of outlier loci were detected with Arlequin than with BayeScan (Table 17). Consistent outliers were found with all analyses (i.e. outliers detected with all statistical approaches). When using samples from both slopes, a big difference was obtained in the number of outliers identified between BayeScan and Arlequin approaches. Further, the proportion of outliers under divergent and balancing selection showed opposite proportions between both approaches, being balancing outliers more frequent with Bayescan (86.4%) while divergent with Arlequin (92.9%), this pattern being even more accentuated in the hierarchical scenario. Some SNPs were shared between approaches for within each type of selection, but never between different types of selection. Balancing selection would be *a priori* more easily detected using F_{ST} outlier tests in a highly divergent genetic scenario, such as that occuring between Atlantic and Mediterranean slopes. Most consistent outliers detected here showed balancing selection (22 of 26 consistent outliers). It is likely that a large part of the divergent outliers detected with Arlequin may be false positives. Three SNP pairs related to divergent selection were found within a 500 kb window in the brown trout genome, and gene mining was performed to identify candidate genes and their associated GO terms (Table 18).

Table 17. Outlier loci detected using different subsets of samples and models. In bold type, the number of outlier loci under divergent selection, in italic type number of outlier loci under balancing selection. MI: Miño Basin; DU: Duero Basin; MED (Ter Basin): Mediterranean Basin. With the first subset of samples two Arlequin hierarchical analyses were performed: (A) hierarchical groups according to slopes (Atlantic and Mediterranean slopes) and (B) hierarchical groups according to river basins (i.e. Miño, Duero and Ter).

ID	Subset of samples	SNPs	BayeScan (*q*-value < 0.05)	Arlequin (*p*-value < 0.01)	Arlequin hierarchical (*p*-value < 0.01)		Consistent outliers (detected in all analyses)
1	MI+DU+MED (2004)	20,293	184 (**25**;*159*)	3,114 (**2,267**;*847*)	(A) 2,466 (**2,433**;*33*)	(B) 2,256 (**2,185**;*71*)	26 (**4**;*22*)
2	Atlantic Slope	15,158	217 (**96**;*121*)	461 (**379**;*82*)	311 (**256**;*55*)		84 (**44**;*40*)
3	Duero Basin	10,777	95 (**47**;*48*)	249 (**202**;*47*)	–		67 (**38**;*29*)
4	Miño Basin	11,332	13 (**13**;*0*)	608 (**263**;*345*)	–		4 (**4**;*0*)

Table 18. Gene mining in the brown trout genome on the three windows (500 kb), where pairs of outlier loci were detected when comparing all samples of Mediterranean and Atlantic slopes: (1) chromosome 4 between 32,491,083 and 32,991,083 bp (light yellow); (2) chromosome 21 between 1,125,412 and 1,625,412 bp (light blue); and (3) chromosome 22 between 2,163,580 and 2,663,580 bp (light green).

ID	GENE DESCRIPTION	GO TERMS FOR BIOLOGICAL PROCESS (LEVEL 3)
pde4ba	cAMP-specific 3',5'-cyclic phosphodiesterase 4B-like	**GO:0006807** nitrogen compound metabolic process **GO:0071704** organic substance metabolic process **GO:0050794** regulation of cellular process
sgip1a	SH3-containing GRB2-like protein 3-interacting protein 1	**GO:0071840** cellular component organization or biogenesis **GO:0051234** establishment of localization
ENSSTUG00000015164	relaxin-3-like	**GO:0050794** regulation of cellular process **GO:0051716** cellular response to stimulus **GO:0007154** cell communication
ENSSTUG00000015172	Possibly ORF2p gene in *Danio rerio*	
ttc4	tetratricopeptide repeat protein 4-like	**GO:0007275** multicellular organism development **GO:0048856** anatomical structure development **GO:0009653** anatomical structure morphogenesis
lurap1	leucine rich adaptor protein 1-like	**GO:0007275** multicellular organism development **GO:0007165** signal transduction **GO:0051716** cellular response to stimulus
ENSSTUG00000015254	Possibly col24a1 gene in *Danio rerio*	

GO terms level 3 for Biological Process category were obtained with Blast2GO.

Table 18. ***(Cont.)***

ID	GENE DESCRIPTION	GO TERMS FOR BIOLOGICAL PROCESS (LEVEL 3)
ankhb	ANKH inorganic pyrophosphate transport regulator	**GO:0055085** transmembrane transport **GO:0051234** establishment of localization
RNF170	ring finger protein 170	**GO:0044238** primary metabolic process **GO:0006807** nitrogen compound metabolic process **GO:0071704** organic substance metabolic process
lrrc6	leucine rich repeat containing 6	**GO:0006928** movement of cell or subcellular component **GO:0007275** multicellular organism development **GO:0009653** anatomical structure morphogenesis
kcnq3	potassium voltage-gated channel subfamily Q member 3	**GO:0055085** transmembrane transport **GO:0050794** regulation of cellular process **GO:0032879** regulation of localization
efr3a	EFR3 homolog A	**GO:0051641** cellular localization **GO:0033036** macromolecule localization
adgrl2a	adhesion G protein-coupled receptor L2a	**GO:0007165** signal transduction **GO:0007154** cell communication **GO:0007275** multicellular organism development

GO terms level 3 for Biological Process category were obtained with Blast2GO.

The highest number of consistent outliers was found in the whole Atlantic Slope followed by within Duero Basin. The total number of consistent outliers were 140, some of them appeared as consistent outliers in different analyses (Table S9). These outliers were located on 37 out of 40 chromosomes of the brown trout karyotype. A total of 63 outliers were in intergenic regions (IGR), 76 within genes (65 outliers in introns, 10 in exons and one in 3'-UTR), and one in a pseudogene (Table S10). Four of them were placed in overlapping genes. Some consistent SNPs were in genes involved in important functions such as immune response (*arhgef2*, *rnf8*) or oxygen transport (*hbaa2* and *hbba2*) (Discussion and Table S10).

Population structure inference conducted by outlier SNPs

The population structure analyses were performed with the two subsets of samples with the higher number of consistent SNPs detected under divergent selection, i.e. Atlantic Slope and Duero Basin populations (Table 17). Within the Atlantic Slope the K estimators showed different values: ΔK (2), MedMean K (5) and Mean LnP(K) (9). This disparity between the different estimators happened in the previous analysis with the Atlantic Slope (see Fig. 17). With K = 2, two clusters were delineated corresponding to Miño Basin plus two populations from Duero Basin (AG1 and BL) and to the inner part of Duero Basin, respectively (Fig. 23). Higher Ks determined the breaking down of some populations from the two initial clusters (VI, AG1, BL), as occurred with the whole SNP panel analysis (Fig. 17). The individual AG1-22, detected as a migrant as with the whole dataset, was also identified here. Nevertheless, the inner Duero Basin populations were more consistently grouped (Fig. 23) than with all SNPs where the CE and OM populations showed to a certain degree a specific component (Fig. 19).

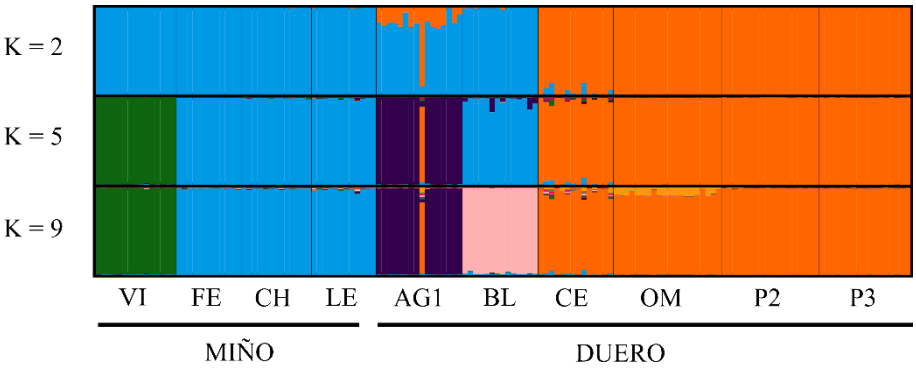

Figure 23. CLUMPAK plot of STRUCTURE assignment with different Ks, (K = 2, 5 and 9). N = 151 wild individuals from 10 populations of the Atlantic slope. A total of 44 consistent outlier SNPs under divergent selection were used. Each individual is represented as a vertical bar partitioned into segments according to the proportion of the genome belonging to each of the clusters identified (K) by STRUCTURE.

Within Duero Basin the most likely number of populations units was similar among K estimators: MedMean K, ΔK, and Mean LnP(K) were two, three and four, respectively. By increasing the K values there was more information about populations structure (Fig. 24).

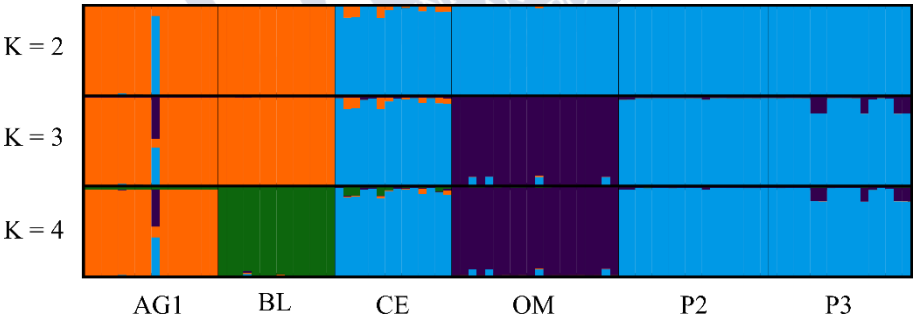

Figure 24. CLUMPAK plot of STRUCTURE assignment from K = 2 to 4. N = 99 wild individuals from six populations of Duero Basin. A total of 38 consistent outlier SNPs under divergent selection were used. Each individual is represented as a vertical bar partitioned into segments according to the proportion of the genome belonging to each of the clusters identified (K) by STRUCTURE.

With K = 2, two clusters were delineated corresponding to the lower-course (AG1-BL) and the inner Duero Basin. Higher Ks

determined the breaking down of some populations from the initial clusters (BL and OM). CE populations was in the same cluster as Pisuerga populations (P2-P3) with all tested Ks.

For DAPC analyses, following the cross-validation method, the number of PCs recommended retained at least 90% of variance. For Atlantic Slope, the lowest BIC value (187.86) corresponded to K = 9, despite different Ks showed very close values. After the cross-validation method was applied, 22 PC axes and four discriminant functions were retained (95.7% of the variance). VI appeared clearly separated from the rest of Miño Basin populations (Fig. 25). Inner Duero populations appeared as a single group in both scatterplots. In Fig. 25(A) the grouping was similar to Fig. 23 (K = 5 and lower K values, plots not showed). In Fig. 25(B) the migrant individual in AG1 population is clearly displayed.

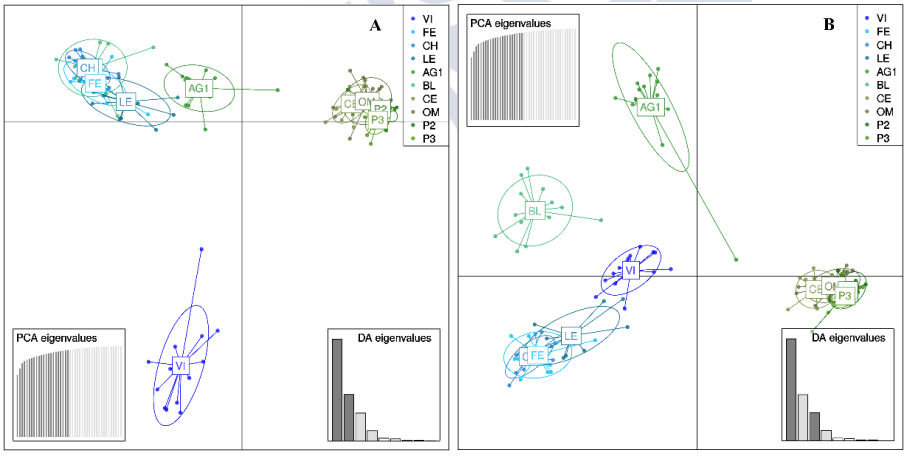

Figure 25. Scatterplots of individuals on the three principal DA eigenvalues of DAPC. Populations are coloured according to their basin, blue shades for Miño and green shades for Duero. The graph represents the individuals as dots and the groups as inertia ellipses. PCA and DA eigenvalues are displayed inset. A: Scatterplot with the two first DA eigenvalues. B: Scatterplot with the first and third DA eigenvalues.

For Duero Basin analyses the lowest BIC value (44.37) corresponded to K = 9, despite different Ks showed very close values.

After the cross-validation method was applied, 12 PC axes and three discriminant functions were retained (96.2 % of the variance). In both scatterplots BL is separated from the rest of the populations and AG1-22 sample appeared near the inner Duero populations grouping (Fig. 26), as in the previous scatterplots. In Fig. 26(B) OM appeared separated from inner Duero populations, as in Fig. 24 with Ks > 2.

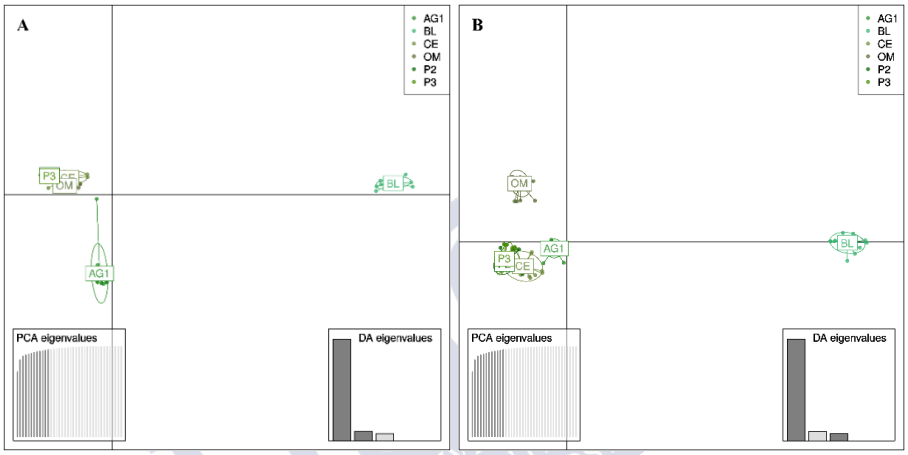

Figure 26. Scatterplots of individuals on the three principal DA eigenvalues of DAPC. All populations belong to Duero Basin. The graph represents the individuals as dots and the groups as inertia ellipses. PCAs and DAs eigenvalues are displayed inset. A: Scatterplot with the two first DA eigenvalues. B: Scatterplot with the first and third DA eigenvalues.

5 DISCUSSION

5.1 IMPACT OF DIFFERENT SNP PANELS FROM TWO BUILDING-LOCI PIPELINES ON POPULATION GENOMIC INFERENCES

In the last decade, the binomial NGS / RAD-seq has been the choice for genomic screening in many studies due to the vast number of genetic markers identified and genotyped in a single step. In this context, species with low genomic information have been targeted for population genomics and evolutionary studies broadening the opportunities for more refined approaches regarding conservation genetics and breeding programs. Nevertheless, the effect of genomic architecture, genetic diversity, and population structure into the outcomes of these techniques (number of SNPs, genotyping confidence) in the target species are essential issues to be addressed using both with simulation and real data approaches. These issues are not only important for the wet-lab protocols, but also for the bioinformatic pipelines to be used to analyse the huge amount of data produced. Technical decisions on the reduced representation method and restriction enzymes selection to be applied when constructing libraries are critical. When a reference genome is available the number of potential loci obtained with different restriction enzymes (e.g. using ExtractSites.pl https://github.com/Eli-Meyer/2bRAD_utilities) and the uniqueness of RAD-loci should be tested (e.g. using EvalFrags.pl from

2bRAD_utilities). For instance, in Manila clam was predicted that the percentage of the genome constituted by repetitive elements and combined transposable elements could exceed 70% (Yan *et al.* 2019). This genomic information should be considered to make the best technical decisions. Without reference genome, different REs can be tested if the budget allows it (see Box 1 in Barbanti *et al.* 2020), to improve the percentage of reads to build up confident loci. In the same way, the different performances of software and bioinfomatic pipelines might depend on the species and its genomics context. These can affect not only the number of markers found, but more importantly, the biological conclusions drawn. A repertoire of bioinformatic publications to manage the large amount of genomic data at different stages (e.g. building-loci pipelines, SNP filtering steps) has been published to serve as guidelines for researchers with limited experience in the field and advices for bioinformatic "Gordian Knots".

The panels used in this study came from species that differ in their genomic architecture, polymorphism and population structure, and these factors could influence the results obtained also depending on the different population parameters used (e.g. global F_{ST}, allelic richness). Nevertheless, the results obtained in our study within species using the four *de novo* panels (i.e. STA, ALT, COM, MER) were roughly similar for all the population parameters evaluated. Accordingly, biological inferences would hardly change. Minor differences were found in the number of suggestive outliers detected in common cockle. In this case, the number of suggestive outliers detected could be related to the total number of SNPs of each panel. But beyond this observation, our results suggest that whatever the pipeline chosen similar results are obtained with a *de novo* approach. The number of initial and final SNPs obtained with reference genome was lower than obtained with a *de novo* approach. This would be related to the high number of input reads removed by duplicate or multiple alignment to the reference genomes

due to the short length of 2b-RAD reads before going to the building-loci pipeline. Some population parameters between both approaches showed relevant differences (see F_{IS} values in brown trout; Table S5).

A practical approach to decide between pipelines with different building-loci strategies for handling genotyping by sequencing (GBS) data is to assay trials with a small subset of data and check for their results using a meaningful set of population parameters, previously selected according to the objectives of the study. Indeed, using the number of SNPs obtained as the main criterion (Puritz *et al.* 2014) to decide the best building-loci pipeline to be used is not advisable, since a higher number of SNPs does not necessarily indicate a better stacking and confident RAD-seq data (Díaz-Arce *et al.* 2019), and consequently, it might have a negative impact on the confidence of results and biological inferences. The initial number of SNPs obtained with STA and ALT pipelines across the different species tested was rather similar except for the brown trout and the small-spotted catshark. These species showed the lowest median coverage, near to the selected threshold coverage filter (8x) hence the differences on the number of putative loci from input data. While Stacks 2, with a *de novo* approach, starts with individual data demanding a number of identical reads to build a locus (see Material and Methods), Meyer's 2b-RAD v2.1 works with a combined subset of confident reads from all samples to build a global reference panel to which align every read. In cases with low coverage, a less demanding criterion to build loci can produce large differences in the initial number of SNPs. Nevertheless, through the filtering steps, the SNP number from building-loci pipelines converged in both species and importantly, the population parameters between SNPs panels were similar. Once chosen the pipeline, it would be recommendable to run several trials with different parameters to properly adjust them to the dataset. For instance, -M in Stacks (which defines the maximum nucleotide differences allowed between intraindividual putative loci)

depends on the levels of polymorphism of the species and -m in Stacks on the existing coverage (Paris *et al.* 2017). In the same way as for the choice of the building-loci pipeline, it would be advisable to choose parameters taking into account the results from population outcomes, since there is not a unique pipeline suited to every situation, as already indicated Torkamaneh *et al.* (2016).

After the building-loci pipeline, it is important to adjust filtering (criteria and order; O'Leary *et al.* 2018) according to the particular scenario of each species (e.g. sequencing and genotyping errors, duplicated loci; Benestan *et al.* 2016). Since the filtering parameters are dataset dependent (Hendricks *et al.* 2018), the filtering criteria should be adjusted accordingly (e.g. the stringency of MAC filtering step is sample size dependent). For instance, the number of SNPs was markedly reduced through filtering steps and the highest difference in the percentage of retained SNPs was found among species. In the study by O'Leary *et al.* (2018) the percentage of retained SNPs ranged from 0 to 63% using the same filtering pipeline with four marine fish species. In our study, the three SNP/RAD-locus filter used to avoid inconsistent RAD-loci could not work well for highly polymorphic species or taxa (e.g. bivalves). Furthermore, the POP filter (i.e. 60% call rate per population) could be applied not so stringently since in previous studies qualitative interpretations of population parameters were maintained in most cases (Shafer *et al.* 2017; Wright *et al.* 2019), and sometimes even improved (Hodel *et al.* 2017). Notwithstanding, the drawback could be using larger SNP panels for similar information. If well, for some studies it is fundamental to achieve the highest density of SNPs possible (e.g. linkage disequilibrium, outlier detection and gene mining, Genome-wide association study; GWAS). The biggest difference between pipelines was found with the MAC filtering step in brown trout which could be explained by the higher average of missing genotypes per SNP (MAC is sample size dependant) and the lower coverage per

RAD-locus (misclassification of heterozygotes) in the ALT pipeline. Finally, more filtering steps might be necessary, especially when working without a reference genome (e.g. F_{IS} SNP filtering step when paralogs or null alleles can be a problem to avoid misinterpretations); this is the case of HWE deviations in brown trout caused by potential paralogs whose impact can be reduced using a reference genome.

Attention should be paid to the order of the different filtering steps because this can alter the final SNP panel. When adjusting the filtering parameters, it would be advisable to consider not exclusively the number of removed SNPs at each step separately, since they could result from the interaction among filtering steps. For instance, the coverage filter determines the increase of missing data which influences the percentage of SNPs eliminated by MAC and population representation filters, according to the stringency of the coverage threshold used. Furthermore, missing data may be due to a lower coverage than the selected threshold or for not being genotyped by the building-loci software with the genotyping options selected (e.g. previously selected nucleotide frequencies range to genotype in ALT pipeline). We found that the last could be the main source of COM SNPs genotyping differences between both building-loci pipelines excluding Manila clam. This means that the ALT pipeline genotyping parameter should be improved by choosing appropriate ranges for each species. The objective of any filtering strategy is removing SNPs that are not reliable without losing informative SNPs. Different factors can influence the filtering criteria, e.g. to achieve the number of SNPs required to meet the research goals. In this sense, a panel made up with markers found by two different pipelines should ensure reliability. It was found that 67% of SNPs from Stacks panel were common with UNEAK panel, using a *de novo* approach in soybean (*Glycine max* L.) data (Torkamaneh *et al.* 2016). With reference genome the overlap percentages among Stacks and other building-loci pipelines ranged

from 76 to 96% (Torkamaneh *et al.* 2016). Using a reference genome approach the percentages of shared SNPs between Stacks with SAMtools and GATK ranged from 7.3 to 71.4% (Wright *et al.* 2019). The lowest values could be partially explained because Stacks panel recruited many more SNPs than the other building-loci pipeline. The lowest percentage of COM SNPs taking STA panel as genotyping reference in our study (i.e. 23.9% in small-spotted catshark and 43.0% in Manila clam) were in panels with less than 1,000 SNPs. These low values may be explained by a strong filtering effect, on shared SNPs between pipelines. The highest number of COM SNPs were detected when STA panels included the highest number of SNPs, around 74% in brown trout and 81% in silver catfish. Despite including a lower number of SNPs, the COM panels provided roughly similar results to the larger ones. This suggests that most informative markers are retained downstream, with the advantage of working with a reduced panel that can simplify and speed-up analyses. In the study by Díaz-Arce *et al.* (2019) the possible effect of SNP number on F_{ST} estimation using reduced SNPs subsets was tested and similar values regarding the full panel were obtained. Moreover, estimated genotyping accuracy may be higher with SNPs shared by more than one building-loci pipeline according to Torkamaneh *et al.* 2016. The impact of genotypic differences between shared SNP panels was low, such as those obtained by Wright *et al.* (2019).

Summarizing, the results obtained suggest that both building-loci pipelines are adequate and provide more confident results adjusting parameters and SNP filtering steps to the research context. Despite the differences observed in the number of SNPs among *de novo* approach panels, this seems not to affect dramatically the conclusions, at least in the biological scenarios managed in this study. When there is no reference genome, a COM panel could be interesting in terms of SNP panel consistency with species with high genomic complexity. In a

general way for some population parameters, to have less SNPs do not imply loss of biological information and a COM panel could increase data reliability in these cases. In the case of choosing this option differences in genotyping between pipelines should be checked, although in this study genotyping differences between pipelines were infrequent. The main source of genotyping differences in COM SNP panels was missing data and had different sources. On one hand, different genotypes can be obtained due to the different building-loci pipeline parameters to call genotypes (e.g. --alpha at Stacks) and the different alignment strategies (e.g. reads with alternative allele can be stacked into another putative loci). For missing genotype differences, it should be considered that even RAD-loci showing high coverage, might have missing data if building-loci pipeline parameters involved in genotyping are not properly set up. Furthermore, small differences in the building-loci pipeline could have more influence in the number of missing genotypes when working with low coverage loci. Anyway, it would be advisable to use a few intra-library and inter-library sample-replicates to estimate genotyping errors (Mastretta-Yanes *et al.* 2015) to increase the confidence in our data, especially if the RAD-seq libraries are designed with low estimated coverage per locus (e.g. around 10x), due to the impact of coverage in genotyping error rates (Fountain *et al.* 2016).

5.2 BROWN TROUT POPULATION GENOMICS

Impact of restocking in natural populations

River drainages of Spain have been massively restocked to counterbalance population depletion since the 70's using a hatchery stock of Central-European origin. This original stock was distributed among all hatcheries across Spanish geography since then, and consequently a slight genetic differentiation can be detected nowadays within an essentially homogeneous genetic background ($F_{ST} < 0.05$;

Almodóvar *et al.* 2006; García-Marín *et al.* 2018). Anyway, the genetic divergence between the hatchery stock and Spanish brown trout wild populations is much larger than among hatchery stocks used for restocking (García-Marín *et al.* 1991; Martínez *et al.* 1993; Vera *et al.* 2013), which supports the use of a single hatchery (BA14) as reference in our study to check the incidence of restocking in Atlantic and Mediterranean drainages.

The impact of restocking has been evaluated until recently using the nuclear diagnostic marker *LDH-C**, fixed for the *90 allele in hatchery stocks and for the *100 allele in wild populations (Morán *et al.* 1991). This method showed important limitations in scenarios where a moderate or high ancient introgression occurred because of the viability of hybrids and their offspring. In this case, the use of *LDH-C** could be inaccurate at population level, and not useful at individual level since *LDH-C**100/100 individuals could have an important hatchery genomic background and viceversa. Estimates of introgression based on mtDNA markers have been used as well, nevertheless hatchery mtDNA haplotypes and hatchery nuclear markers do not necessarily match either at individual (Sanz *et al.* 2006) or at population level (see Table 6 in *Plan de gestión de la trucha común en Castilla la Mancha 2019*). The use of a high number of SNPs distributed across the whole brown trout genome and the availability of a reference hatchery sample from Bagà, enabled us a more accurate classification of individuals by hatchery ancestry using the probability of membership assignment with STRUCTURE, as previously reported (Hansen *et al.* 2001; Prado *et al.* 2018). We applied a conservative *q* threshold to assign individuals as wild (q_W) or hatchery (q_H) ancestry by applying a self-assignment test in a population of known genomic background (Bagà). A slightly higher conservative threshold than that used in other investigations on brown trout (Sanz *et al.* 2009) and turbot

(*Scophthalmus maximus*; Prado *et al.* 2018) was used in accordance with our population scenario.

The incidence of restocking detected in our study was variable across the populations studied in the Mediterranean and Atlantic slopes. Although a higher impact was observed in the Mediterranean drainage than in the Atlantic one, as previously reported (Almodóvar *et al.* 2006), the differences among Mediterranean populations were remarkable, sometimes between populations separated by a few kilometres. Temporal replicates from Ter River (TE) did not show any hatchery incidence, while samples from Queralbs, pertaining to the same river drainage (Ter Basin), were the most affected. A similar observation was reported by Araguas *et al.* (2017) using five microsatellites and the diagnostic locus *LDH-C**. Núria and Queralbs locations are separated by five kilometres in the same basin, but currently could be partially isolated by different barriers (e.g. dam in Daió hydroelectric power plant). Such barriers might favour the genetic integrity of the brown trout from Ter River. Dams can act as a barrier for alien species invasion and in some cases, they have been built to protect native species (Dana *et al.* 2011). In the Atlantic drainage, the same individuals from BL were removed by in Martínez *et al.* (2007) being considered as pure hatchery (BL-15) and F1 (BL-24) using the *LDH-C** locus, while here a more refined genomic constitution of those individuals was achieved. Usually, to classify individuals as wild or hatchery introgressed to take management decisions, a global ancestry approach using tools as STRUCTURE was employed. Nevertheless, local ancestry inference would be more powerful to detect small chromosomic regions affected by hatchery introgression. As in the study of Leitwein *et al.* (2018), some individuals previously identified as "pure wild" showed "introgression signals" with local ancestry approach. Chromosomic regions that may be prone or resistant to introgression were detected in our study. However, to avoid interpreting sampling effects as

introgression patterns, more admixed samples should be analysed and compared to establish consistent resistant or prone regions to introgression.

We intended to develop a more powerful, informative and cheaper molecular tool to evaluate restocking in Spanish drainages by using different panels of progressively higher power to detect hatchery ancestry individuals. We assumed that the most consistent *qH* values were those obtained with the whole SNP dataset, thus being used as reference for evaluating the performance of the different subsets. However, low informative SNPs might be filtered according to F_{ST} values defining subsets of high-resolution SNPs. So, we identified the most informative markers using F_{ST} between wild and hatchery samples and detected nine diagnostic SNPs capable to identify most hatchery ancestry individuals in our sample. A second and a third subset of 38 (F_{ST} > 0.99) and 214 (F_{ST} > 0.95) SNPs were also evaluated for the higher resolution and the lower cost as possible. The performance of the three SNP subsets for the classification of individuals according to hatchery ancestry using, either the STRUCTURE approach or the proportion of hatchery alleles at individual level using diagnostic loci, was remarkable and the correlations obtained with the whole genomic data were highly significant for the three subsets (*p*-value < 0.001). Some discrepancies at population and individual level were related to individuals with a slight hatchery ancestry, especially in CE population. Further, the hybridization and introgression across the genome is not homogeneous (see for instance Wang *et al.* 2019 for soybean) and a small number of AIMs (Ancestry Informative Markers) may not be enough. These SNPs, identified *in silico* from RAD-seq, should be validated in the future with techniques well fitted to handle small SNP panels (e.g. SNaPshot, Sequenom, TaqMan) to devise the best molecular tool combining statistical power and low price. The current price of RE-digestion for *LDH-C** genotyping is close to 10€ /

individual, while a single multiplex for Sequenom could even be cheaper. Similar molecular tools have been successfully applied for the identification of hybrids in aquatic organisms (Maroso *et al.* 2018, 2019). Nevertheless, these SNPs should be tested with new individuals and more samples covering a wider distribution of the Iberian Peninsula to evaluate their performance at individual and population level regarding the previous results obtained with the *LDH-C** locus.

Genetic diversity

Preservation of genetic diversity is a key point to maintain the potential for adaptation of natural populations to environmental changes (Frankham *et al.* 2010), which are rapidly affected by human activities and the ongoing climate change. Since the 90's until recently, the most commonly genetic markers used to estimate genetic diversity for conservation and management of bioresources were microsatellites (Saint-Pé *et al.* 2019). A lot of studies have been performed with microsatellite markers even during the last 10 years (Araguas *et al.* 2017; Berrebi *et al.* 2019; Vera *et al.* 2013, 2018; Vilas *et al.* 2010). The arrival of NGS techniques and related techniques have allowed a quick identification and cheap genotyping of thousands of SNPs, even without reference genomes, the so-called genotyping by sequencing (GBS) techniques (Davey *et al.* 2011; Robledo *et al.* 2018). The arrival of SNPs has let population and individual screening of genomes for more accurate estimation of genetic diversity and structure, and especially, the identification of footprints of selection (Bernatchez 2016). The fact that SNPs are usually biallelic markers while microsatellites are hypervariable with tens of alleles per locus, makes that genetic diversity estimators cannot be directly comparable within populations (Bouza *et al.* 2001; Saint-Pé *et al.* 2019; Vilas *et al.* 2010), but still the relative genetic diversity among populations can be compared with both types of markers.

The average figures of genetic diversity found in the present study for the Iberian Peninsula were always much lower than those reported with microsatellites (Bouza *et al.* 2001; Martínez *et al.* 2007) for all populations (H_e = 0.064 ± 0.034), with Galician populations (0.106 ± 0.020) showing higher diversity than Duero (0.045 ± 0.024) and Mediterranean populations (0.040 ± 0.004). Low genetic diversity in Mediterranean Slope is expected due the important impact of genetic drift due to unstable hydrology and isolation of river basins (Araguas *et al.* 2017; Vera *et al.* 2013). Conversely, populations above the parallel 42º N, including the Galician region, are expected to show higher genetic diversity due to stable hydrology and inter-basin connection through the anadromous trout, the so-called sea trout (Antunes *et al.* 2006; Bouza *et al.* 1999; García-Marín *et al.* 2018; Östergren and Nilsson 2012). The migration of sea trout would explain a higher effective population size (N_e), and accordingly, higher genetic diversity. Our results, therefore, meet to previous observations. However, comparison with other studies in northern regions have shown much higher genetic diversity than that found here. For instance, a wide survey across Northern European populations of *S. trutta* (72 locations placed in Great Britain, Germany, and Scandinavia) with 3,872 SNPs showed an average H_e > 0.30 (Bekkevold *et al.* 2020), three times higher than that found in Galicia in our study. These differences found among populations of the same species with SNPs has usually to do with the filtering performed in each study. Thus, the SNPs used by Bekkevold *et al.* (2020) were genotyped with a SNP-chip where highly polymorphic SNPs had been selected, so rendering an average MAF of 0.28 and only 1.75% showing MAF < 0.05; this unavoidably upwards H_e estimations. Similar genetic diversity to that found by Bekkevold *et al.* (2020) was obtained in Southern Baltic area by Bernaś *et al.* (2020) with 3,843 SNPs in a region where sea trout also occurs. Average MAF in our study was 0.13, with a much higher proportion of loci with MAF < 0.05 (44.0%). Many of these alleles were private of specific

populations both due to the high structure of brown trout (Ferguson 1989), but also to the rather low sample size here managed. Another recent study in Finnish populations with 4,876 SNPs showed lower genetic diversity, with H_e values around 0.1 (Lemopoulos *et al.* 2019a), very similar to our results. Again, the more relaxed filtering chosen by these authors (MAF = 0.005) underlies this estimation. Caution should be taken when comparing genetic diversity among different studies on the same species, particularly paying attention to the filtering process followed to obtain the SNP panel.

Hatchery populations used for restocking have shown higher genetic diversity than wild ones because of their mixed genetic origin to enhance genetic diversity. Accordingly, most studies in Spain have supported higher diversity in the hatchery stock than in wild populations. Araguas *et al.* (2017) showed higher genetic diversity in a hatchery sample (BA14) with microsatellites than in wild Mediterranean populations, such as been shown here, with the only exception of NU14. However, unlike previous studies with allozymes and microsatellites that had shown lower genetic diversity in Galician populations than in the hatchery stocks (Bouza *et al.* 1999; Martínez *et al.* 1993), FE and CH populations belonging to Miño Basin showed higher SNP diversity than the Bagà hatchery stock. A possible explanation is that these populations are in an area where hybridization between mtDNA lineages has been reported (Bouza *et al.* 2008; Vilas *et al.* 2010). Higher values in admixed populations are expected when the genetic population units involved have very different allelic frequencies or even fixed allelic variants. Studies in other countries across Europe have also supported the higher diversity of hatchery stocks (Berrebi *et al.* 2019; Bohling *et al.* 2016; Lemopoulos *et al.* 2019; Martínez *et al.* 2007). Nevertheless, as explained in the next section, in the putative hybrid populations from Duero this effect was not so clear. BL had the highest genetic diversity value in Duero Basin

and P2 the lowest. This pattern was the same observed previously with microsatellites (see Appendix in Martínez *et al.* 2007), but CE and OM showed similar values to the other populations. Almost all wild populations had slightly negative values of F_{IS}. Notwithstanding, in Vilas *et al.* (2010) and Araguas *et al.* (2017) most F_{IS} values were slightly positive (i.e. heterozygote deficit), possibly due to a higher incidence of null alleles with microsatellite data. Finally, all populations were in Hardy-Weinberg equilibrium. This result suggests that hatchery allelic variants introduced during decades would be integrated into the reproductive pool of wild locations of the studied regions. However, except for NU04, the percentage of loci in HW disequilibrium (i.e. *p*-value <0.05) increased when individuals of hatchery ancestry were included.

Genetic differentiation and structure

The genetic differentiation among populations of Iberian Peninsula was from moderate to very high, reflecting the high structuring of brown trout in this area, in accordance with the high population structure characterising this species (Ferguson 1989). This high population differentiation was reported in previous surveys with different genetic markers including allozymes, microsatellites and mtDNA (see Table 4.1 in García-Marín *et al.* 2018). However, comparisons with different genetic markers should be made with caution due to the different evolutionary forces driving their divergence pattern, as previously reported (see Table 6 in Ősz *et al.* 2018). In addition, the filtering pipeline followed in the case of SNPs is critical, as outlined before for the total amount of genetic diversity detected; this could have some influence in its distribution across the hierarchical levels considered. An extreme example is the use of diagnostic SNPs to unravel structure such in Sušnik *et al.* (2015), who reported an F_{ST} = 0.981 between Atlantic and Danube brown trout populations.

The deep structure reported for brown trout is also illustrated by the significant F_{ST} values found between populations in the same watershed (Lemopoulos *et al.* 2019a) or between localities separated by less than a few kilometres in the same stream (Bouza *et al.* 1999; Carlsson and Nilsson 2001; Sanz *et al.* 2019). As expected, in our study the lowest F_{ST} value was obtained between the two samples of the Bagà hatchery strain (F_{ST} = 0.015), representing released individuals and spawners of the same broodstock (BA14 and S, respectively). Also, temporal replicates showed low genetic differentiation, particularly between Ter samples (F_{ST} = 0.017), which indicates reduced fluctuations of gene pools in this population across generations. Nevertheless, the pairwise F_{ST} between Núria replicates was quite higher (F_{ST} = 0.087), likely due to the notable impact of hatchery releases in Núria and Freser populations (e.g. NU14, QB14) which might determine a reduction in the effective native population size with a higher genetic drift than in Ter populations. Ter and Núria sampling locations are within genetic refugees created in 1997 by Autonomous Government of Catalonia (Araguas *et al.* 2009), where restocking was banned although fishing was allowed. Since 1991, there has not been recorded hatchery releases (Araguas *et al.* 2004). Interestingly, the impact of hatchery was higher in Núria and Freser rivers, even higher after the ban of restocking (Araguas *et al.* 2017). This could be related to one naturalised hatchery population located in a tributary of the river Freser (PAR14 population, Araguas *et al.* 2017), likely representing a source of migrants for Núria and Queralbs populations but not for Ter. The low connectivity in the final stretch of the Freser river, before it flows into the Ter River (*Agència Catalana de l'Aigua* 2020) could explain the different impact between Núria and Ter samples. The high differentiation obtained among Duero populations (global F_{ST} = 0.460) is very similar to that reported previously with allozymes (average G_{ST} = 0.455; Bouza *et al.* 2001) and microsatellites (average R_{ST} = 0.413; Martínez *et al.* 2007) suggesting scarce gene flow among populations

caused by habitat fragmentation, likely enhanced in the last decades by the construction of multiple hydroelectric dams. The highest pairwise differentiation within Duero was detected between AG1 and P2 as in the previous reports. Lower global (F_{ST} = 0.192) and pairwise (0.050-0.326) F_{ST} values were found in Miño Basin where hybrid populations showed little differentiation from the reference populations used. The high structure detected among populations belonging to the same basin could mean that each brown trout population could represent a unique gene pool and this should be considered for appropriate management and conservation plans to be developed by administrations.

Different hierarchical hypotheses were tested using the partition of the total variance (AMOVA) according with different criteria involving previous data and our own analyses with STRUCTURE software. A significant proportion of the variance was detected when comparing different slopes, an expected outcome considering the isolation of Atlantic and Mediterranean slopes since thousands of generations ago. Also, differentiation was significant between Duero and Miño basins, a fact also observed in Galicia and northern areas of Spain (Bouza *et al.* 1999; Morán *et al.* 1991), where, unlike Duero, the anadromous form drifts through the sea connecting drainages and diminishing differentiation. As reported by Antunes *et al.* (2006), Duero River would be below the limit of the distribution range of the anadromous form and therefore, gene flow between river basins, if any, would be very residual between the Miño and Duero basins. Nevertheless, statistical significance (*p*-values) is sample size dependant and Fitzpatrick (2009) highlighted that with few populations per group, a significant among-group component (F_{CT}) could be unachievable. This is especially evident in the case of Miño Basin (four populations) and for Duero (six populations), although in this case a marginal significant *p*-value was obtained for some comparisons. The high differentiation between nearby populations observed in brown trout, due to the

homing-behaviour of salmonids and the consequent genetic drift, could be masking the existence of deeper or older structure. Accordingly, the presence of highly differentiated populations is quite usual (Martínez *et al.* 2007) highlighting the need for a greater sampling effort compared to other species or contexts with much lesser structure.

The different statistics used for inferring the number of K in STRUCTURE analyses rendered different values. This might reflect different layers of structure supporting different "true" Ks (Meirmans 2015) and the different performance of these estimators under different scenarios (Puechmaille 2016). Previous studies suggested the presence of hybrid populations in Miño and Duero because of a secondary contact after glaciations between AT and DU lineages segregated in parapatry in those basins, the inner region corresponding to DU and the outlet to AT (Bouza *et al.* 2001, 2008; Martínez *et al.* 2007, 2009; Vilas *et al.* 2010). Depending on the marker used, for instance, the delineation of both lineages and putative hybrid populations was different. So, Bouza *et al.* (2001) using isozymes suggested an inner borderline between both lineages in Duero Basin, while Martínez *et al.* (2007) moved the limit closer to the outlet using microsatellites and mt-DNA RFLPs. However, the hybrid condition of some populations was hard to be confirmed because of the low number of genetic markers used to detect consistent signals of the expected linkage disequilibrium in a hybrid zone. In our study, we found the limit between lineages close to the outlet using thousands of SNPs. In cases like this, where populations genetically most similar to the genotype of one of the parental populations, we talk about a bimodal hybrid zone (Jiggins and Mallet, 2000). Similarly, Martinez *et al.* (2007) suggested a bimodal hybrid pattern with mtDNA RFLPs data in Duero hybrid zones, with a higher proportion of Pisuerga component. Here, the inner populations of Duero might constitute an ancient group concordant with the current high differentiation. Higher levels of structure were also consistent with

different estimators, two of them supporting K = 6, which could be consistent with the impact of the strong genetic drift affecting brown trout populations with effective population sizes below 50. Plausibly, CE and OM could have belonged to a hybridization zone until the construction of the large dams during the last century. Furthermore, a dubious cluster defined as ghost clusters by Guillot *et al.* (2005) was detected, confirmed because it never achieved a membership coefficient level (q) > 0.5 in any individual of this dataset. According to the criteria established by Puechmaille (2016), these ghost clusters could be spurious since never achieves a mean membership coefficient greater than the established threshold value in any population (set to 0.5 by default in StructureSelector).

Effective population size estimation

Effective population sizes obtained would be in the ranges obtained with other populations, using linkage disequilibrium method (N_e estimates from 12 to 285.1 in Esva River; Sanz *et al.* 2019) or temporal methods (see Fig. 2 in Palm *et al.* 2003). The highest values were obtained in Galician populations, where environmental conditions are normally better for brown trout populations (Vera *et al.* 2018). With temporal methods, Ter samples showed higher N_e than Núria samples. This could be caused by the hatchery introgression in Núria, reducing the number of wild individuals in this population, accentuating the variation of allelic frequencies over generations. Much of the N_e estimates showed values below 50. Franklin (1980) suggested that, as a general rule-of-thumb, a N_e = 50 would avoid inbreeding depression in the short-term and N_e = 500 would maintain evolutionary potential in the long-term (i.e. "50/500 rule"), despite these guideline values were revised upwards by Frankham *et al* (2014). It should be noted that these rules are based only on genetic considerations without taking account of demographic stochasticity which in small populations may be fundamental, sometimes lead to population extinction. If we assume

with our data that the rate of random drift in allele frequencies and the rate of decrease in homozygosity are the same (Wang 2016; Ryman *et al.* 2019), there would be several populations that could be extinct, in line with the "50/500 rule", especially if environmental changes that may require adaptation occur (e.g. global warming; Almodóvar *et al.* 2012). All in all, a large part of N_e estimates should be interpreted with caution. According to Jones *et al.* (2016) when the upper bound of an estimate is infinite, the null hypothesis (i.e. the population has the same LD as an infinite-sized ideal population) cannot be rejected with the confidence level chosen. Clearly, this size cannot be real in freshwater species. Infinite N_e can indicate insufficient sample size (Marandel *et al.* 2019). Temporal methods assume that generations are discrete without overlapping, while linkage disequilibrium methods assumes that LD has arisen from genetic drift, and not from other causes as admixture or immigration (Freeland 2011), involving the violation of some assumptions the impossibility of obtaining an N_e estimator with a confident reliability. Higher sample sizes would be necessary to increase the confidence in N_e estimates, enabling subsampling test as recommended England *et al.* (2006). For new analyses, it would be recommendable using higher sample sizes to achieve more confident N_e estimates by linkage disequilibrium method and using temporal replicates to study demographic changes over time and to contrast N_e estimates obtained with both methods.

Adaptive variation

The molecular markers used in previous studies in brown trout from Iberian Peninsula were putative neutral markers (allozymes, microsatellites), not subjected to selective pressures, such as it is also assumed for SNPs. Nevertheless, some genomic regions can be under selective pressure determining a deviation from neutrality on the distribution of genetic diversity across the genome and this information is essential for identifying conservation units for an appropriate

management of genetic resources. Adaptive variation is shaped by environmental factors, either biotic or abiotic, and the correlation between genetic variation and environmental factors would aid to understand the selective pressures on populations in a scenario of global climatic change. These loci departing from neutrality are pointing to genomic regions related to adaptation and usually show lower genetic diversity and higher linkage disequilibrium than the average genome because of selective sweeping. Mining the genome on these regions would aid to identify candidate genes and allelic variants associated to environmental factors and putatively responsible variants relative to adaptive variation. Accordingly, divergent selection would increase genetic differentiation among populations while balancing selection would decrease.

A wide variety of consistent outlier loci in brown trout from Iberian Peninsula were found in genomic regions where genes involved in relevant traits were identified. Nevertheless, even if an outlier is located within a gene, it does not necessarily mean that gene to be under selection since they can be in linkage disequilibrium (LD) with the causative variant in other nearby gene. Outlier loci were detected in genes related to osteogenesis (*plod2*), as well as in kidney and gonad development in fish (*wt1b*, Jiang *et al.* 2017). In haemoglobin genes involved in oxygen transport (*hbaa*/*hbba* genes) showing two electrophoretically different isoHb classes in salmonids (i.e. “anodic” and “cathodic”), which exhibit pronounced differences in O_2 affinity and other properties (Storz 2016), were located a consistent outlier. In genes involved in the response to pathogens such as *arhgef2* and *itpr2*, differentially expressed in head kidney of Atlantic salmon in response to *Piscirickettsia salmonis* (Rozas-Serri *et al.* 2018), and in *rnf8* involved in the virus-host battle in Atlantic salmon (Eslamloo *et al.* 2017), were also detected. Furthermore, one outlier was found within *trpv4* gene, whose expression may be inhibited in the presence of

viruses, allowing individuals to move to warmer regions (behavioural fever; Boltana *et al.* 2018), and another was placed within *wwc1* gene, one of the 57 candidate genes potentially under positive flatworm (*Gyrodactylus salaris*) driven natural selection identified by Zueva *et al.* (2018) in populations of Atlantic salmon (*Salmo salar* L.). Nevertheless, the presence of this parasite has not been confirmed in Spain (Paladini *et al.* 2021). Outliers placed into nutrition-related genes such as *slc6a1* were also detected; expression regulation of this gene has been related with a hedonic eating response to a palatable diet in fish (Díaz-Rúa *et al.* 2020). Despite regions where various outliers are detected should be of special interest, the density of SNPs in our study was not high enough to unravel such associations. Further work in controlled conditions (common-garden experiments) would be recommendable to check the outliers potentially related to target traits. For instance, Lemopoulos *et al.* (2018) identified in brown trout eight outlier SNPs potentially associated with migration tendency. Nevertheless, Lemopoulos *et al.* (2019b) in a common-garden experiment designed for this trait identified nine outlier SNPs different to the previous ones, hypothesizing that migration traits could have a multigenic control. Regrettably, these experiments require budget and time and from a conservation point of view could imply a "paralysis by over-analysis". In brown trout, populations that are relatively close geographically should not be treated as forming part of the same management unit by default. A short-cut could be the bioinformatic identification of a set of outliers SNPs used with whole or neutral marker sets to determine population structure derived from adaptative selection and demography, to define more robust Management Units in a fast way.

6 CONCLUSIONS

In this thesis we have applied a population genomics approach as a tool for decision making in brown trout conservation in the Iberian Peninsula. The bioinformatic processing of data for robust genotyping has been analysed using a diverse panel of aquatic species, i.e. the building of loci and their genotyping from raw sequence data; then, genomic information obtained has been used for study the genetic diversity and structure of brown trout in the Iberian Peninsula.

1. The results here obtained on the diverse case studies analysed show that building-loci pipelines do not have a substantial influence on the estimation of population parameters and derived biological interpretations. The small differences observed between some *de novo* and reference genome derived panels could be solved by improving SNP filtering steps. Anyway, our results cannot be generalised, and users should contextualise building-loci pipelines to their species and population genomics scenarios. One recommendation would be to test building-loci parameters and filtering SNP steps with subset of samples to save hardware resources and computation time. The best parameter set would be those leading to consistent results obtained across different replicates and should be taken using population parameters consistent with the research goals of the study. Despite being time-consuming, this

preliminary step should enhance the robustness of results and biological conclusions by improving the bioinformatic tools applied.

2. Improving the balance between the number of loci necessary to get robust conclusions and available budget is essential to obtain meaningful outcomes depending on the biological scenario. Regardless of the RAD-seq technique used, if good resolution is obtained, the same genomic regions will be covered by SNPs in all samples analysed. The development of a high-density array for brown trout to cover different features under study (e.g. hatchery introgression, lineage identification, selection footprints) in a more standardised way might be interesting in the future to increase SNP density and robust genotyping to facilitate the implementation of conservation measures.

3. A molecular tool to evaluate restocking was designed allowing detection and identification of hatchery ancestry at individual and population levels in admixed populations in a cheaper and more resolution way than the methodology used until now with brown trout populations in the Iberian Peninsula (locus *LDH-C**). With the genomic resolution achieved we could detect small regions of hatchery introgression as well, inferring local ancestry at the chromosome level.

4. All populations affected by hatchery restocking were in Hardy-Weinberg equilibrium, which suggest that hatchery allelic variants introduced during the past decades have integrated and randomized into the genomic pool of wild populations.

5. Large differences in genetic diversity were detected among wild brown trout populations on the Iberian Peninsula. A general pattern of increasing genetic diversity which follows an East-West and South-North gradient was confirmed, with the highest

values in the Miño Basin. This general pattern would be in line with geographical regions with more suitable environmental conditions for brown trout populations.

6. A deep structure was detected between populations belonging to different basins, even between populations within the same river basin. The deepest break detected between Atlantic and Mediterranean slopes would be the consequence of an absence of gene flow due to geographical isolation. Within the Atlantic Slope, Duero and Miño basins barely showed gene flow due to the scarcity of sea trout connecting both basins, if any. Using different subsets of nuclear SNPs, no evident natural intra-basin hybrid zones previously reported were detected. This suggests that only a few traces of the hybridization events because of secondary contacts remains likely because of the strong genetic drift associated to the small effective population size in brown trout.

7. Consistent outliers under selective pressure were detected in the different hierarchical levels considered, mostly under balancing selection between Mediterranean and Atlantic slopes, while under divergent selection in the Atlantic slope both within and between basins. However, we could not estimate the environmental factors driving selection on this genomic region because of a lack of appropriate records in the populations studied. Anyway, suggestive genes associated to those genomic regions under selection would deserve further work, especially if detailed environmental information can be gathered. Analysis of wild populations should be accompanied by experiments to test the confidence of those outliers, such as gene expression analysis in common garden experiments.

The information obtained from the population genomic analysis carried out in this thesis will be invaluable to design conservation

guidelines for brown trout populations of Iberian Peninsula. For instance, different conservation policies could be applied depending on the degree of hatchery ancestry in the brown trout populations: (1) eradication of naturalised hatchery populations that may still exist; and (2) establishment of different fishing quotas (individuals per person per day) according to different degrees of hatchery introgression detected, reducing quotas on rivers with pure genetic wild populations. The detection of natural hybrid zones, not directly induced by human activities, would be interesting from a conservation perspective, to protect natural processes which could increase genetic diversity and viability of populations. Some of the analysed populations showed low effective population size, which could lead to the definition of brown trout populations as vulnerable according to this criterion and subsequently the establishment of conservation measures as temporary fishing moratoria or quota reduction in well-stablished hydrographic units (e.g. streams, river sectors). The high genetic differentiation detected in brown populations belonging to the same river basin, would imply that restocking under conservation principles (use strains genetically similar to the populations restocked) contemplated in conservation plans and laws can be difficult to implement in a strict manner. If implemented, stocks unaffected by past hatchery introgression events should be used, as determined by global and local ancestry approaches. The definition of brown trout refuges, where stocking or fishing should be forbidden, in sections of different rivers with good ecological conditions could be a complementary action for conservation.

Anyway, in order to adopt management and conservation measures, it would be necessary to work with updated information on a series of populations that are representative of different units of the Iberian Peninsula (e.g. rivers, river basins, Management Units). Ideally, strategies should be consensual and shared between different

Administrations for integrated conservation plan beyond the administrative limits. The genomic information obtained, combined with the results from other disciplines (e.g. ecology), should help to make the best management decisions. In any case, no species conservation measure should be addressed without the associated ecological (e.g. conservation status of habitats) and socio-economic contexts (e.g. current legislation, economic and cultural uses).

7 REFERENCES

Adamack AT, Gruber B (2014) PopGenReport: Simplifying basic population genetic analyses in R. Methods Ecol Evol 5:384–387. https://doi.org/10.1111/2041-210X.12158

Agència Catalana de l'Aigua (2020) Informe Connectivitat longitudinal als rius de les Conques internes de Catalunya. Estructures presents als rius i actuacions de millora realitzades. Generalitat de Catalunya: Departament de Territori i Sostenibilitat.

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19:1655–1664. https://doi.org/10.1101/gr.094052.109

Allendorf FW, Thorgaard GH (1984) Tetraploidy and the evolution of salmonid fishes. In: Turner B (ed) Evolutionary genetics of fishes, 1st edn. Plenum Press, New York, pp 1–53

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. Nat Rev Genet 11:697–709. https://doi.org/10.1038/nrg2844

Allendorf FW, Bassham S, Cresko WA, *et al* (2015) Effects of crossovers between homeologs on inheritance and population

genomics in polyploid-derived salmonid fishes. J Hered 106:217–227. https://doi.org/10.1093/jhered/esv015

Almodóvar A, Nicola GG, Elvira B, García-Marín JL (2006) Introgression variability among Iberian brown trout Evolutionary Significant Units: the influence of local management and environmental features. Freshw Biol 51:1175–1187. https://doi.org/10.1111/j.1365-2427.2006.01556.x

Almodóvar A, Nicola GG, Ayllón D, Elvira B (2012) Global warming threatens the persistence of Mediterranean brown trout. Glob Change Biol 18:1549–1560. https://doi.org/10.1111/j.1365-2486.2011.02608.x

Alonso C, Gortázar J, García de Jalón D (2017) Trucha común – *Salmo trutta*. In: Sanz J, Elvira B (eds) Enciclopedia Virtual de los Vertebrados Españoles. Museo Nacional de Ciencias Naturales, Madrid. http://www.vertebradosibericos.org/. Accessed 26 June 2019

Alp A, Kara C, Büyükçapar HM (2003) Reproductive biology of brown trout, *Salmo trutta macrostigma* Dumeril 1858, in a tributary of the Ceyhan River which flows into the eastern Mediterranean Sea. J Appl Ichthyol 19:346–351. https://doi.org/10.1111/j.1439-0426.2003.00455.x

Amoros C, Roux AL (1988) Interaction between waterbodies within the floodplains of large rivers: function and development of connectivity. In: Schreiber KF (ed) Connectivity in Landscape Ecology. Schöeningh, Münster, pp 125–130

Andrews KR, Good JM, Miller MR, *et al* (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet 17:81–92. https://doi.org/doi:10.1038/nrg.2015.28

Andrews S (2010). FastQC: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ Accessed 26 June 2019

Antunes A, Faria R, Johnson WE, *et al* (2006) Life on the edge: The long-term persistence and contrasting spatial genetic structure of distinct brown trout life histories at their ecological limits. J Hered 97:193–205. https://doi.org/10.1093/jhered/esj014

Araguas RM, Sanz N, Pla C, García-Marín JL (2004) Breakdown of the brown trout evolutionary history due to hybridization between native and cultivated fish. J Fish Biol 65:28–37. https://doi.org/10.1111/j.0022-1112.2004.00544.x

Araguas RM, Sanz N, Fernández R, *et al* (2009) Role of genetic refuges in the restoration of native gene pools of brown trout. Conserv Biol 23:871–878. https://doi.org/10.1111/j.1523-1739.2008.01144.x

Araguas RM, Vera M, Aparicio E, *et al* (2017) Current status of the brown trout (*Salmo trutta*) populations within eastern Pyrenees genetic refuges. Ecol Freshw Fish 26:120–132. https://doi.org/10.1111/eff.12260

Araneda C, Lam N, Iturra P (2018) Development and application of sex-linked markers in salmonidae. In: Wang HP,

Piferrer F, Chen SL (eds) Sex control in aquaculture. Wiley, West Sussex, pp 281–296

ARDÁN Galicia (2018) El sector de la Pesca en Galicia. In: Informe económico y de competitividad. Consorcio zona franca de Vigo, pp 407–470

Ari Ş, Arikan M (2016) Next-Generation Sequencing: Advantages, disadvantages, and future. In: Hakeem KR, Tombuloğlu H, Tombuloğlu G (eds) Plant Omics: Trends and applications, Springer, pp 110–135

Arias J, Sánchez L, Martínez P (1995) Low stocking incidence in brown trout populations from northwestern Spain monitored by *LDH-5** diagnostic marker. J Fish Biol 47:170–176. https://doi.org/10.1111/j.1095-8649.1995.tb06053.x

Arlinghaus R, Mehner T, Cowx IG (2002) Reconciling traditional inland fisheries management and sustainability in industrialized countries, with emphasis on Europe. Fish Fish 3:261–316. https://doi.org/10.1046/j.1467-2979.2002.00102.x

Baird NA, Etter PD, Atwood TS, *et al* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3:e3376. https://doi.org/10.1371/journal.pone.0003376

Barbanti A, Torrado H, Macpherson E, *et al* (2020) Helping decision making for reliable and cost-effective 2b-RAD sequencing and genotyping analyses in non-model species. Mol Ecol Resour 20:795–806. https://doi.org/10.1111/1755-0998.13144

Barlaup BT, Gabrielsen SE, Skoglund H, Wiers T (2008) Addition of spawning gravel—a means to restore spawning habitat of Atlantic salmon (*Salmo salar* L.), and anadromous and resident brown trout (*Salmo Trutta* L.) in regulated rivers. River Res Applic 24:543–550. https://doi.org/10.1002/rra.1127

Bekkevold D, Höjesjö J, Nielsen EE, *et al* (2020) Northern European *Salmo trutta* (L.) populations are genetically divergent across geographical regions and environmental gradients. Evol Appl 13:400–416. https://doi.org/10.1111/eva.12877

Benestan LM, Ferchaud AL, Hohenlohe PA, *et al* (2016) Conservation genomics of natural and managed populations: Building a conceptual and practical framework. Mol Ecol 25:2967–2977. https://doi.org/10.1111/mec.13647

Bernaś R, Burzyński A, Dębowski P, *et al* (2014) Genetic diversity within sea trout population from an intensively stocked southern Baltic river, based on microsatellite DNA analysis. Fish Manag Ecol 21:398-409. https://doi.org/10.1111/fme.12090

Bernaś R, Poćwierz-Kotus A, Árnyasi M, *et al* (2020) Genetic differentiation in hatchery and stocked populations of sea trout in the Southern Baltic: Selection evidence at SNP loci. Genes 11:184. https://doi.org/10.3390/genes11020184

Bernatchez L, Guyomard R, Bonhomme F (1992) DNA sequence variation of the mitochondrial control region among geographically and morphologically remote European brown trout *Salmo trutta* populations. Mol Ecol 1:161–173. https://doi.org/10.1111/j.1365-294X.1992.tb00172.x

Bernatchez L (2016) On the maintenance of genetic variation and adaptation to environmental change: considerations from population genomics in fishes. J Fish Biol 89:2519–2556. https://doi.org/10.1111/jfb.13145

Berrebi P, Caputo Barucchi V, Splendiani A, *et al* (2019) Brown trout (*Salmo trutta* L.) high genetic diversity around the Tyrrhenian Sea as revealed by nuclear and mitochondrial markers. Hydrobiologia 826:209–231. https://doi.org/10.1007/s10750-018-3734-5

Berthelot C, Brunet F, Chalopin D, *et al* (2014) The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat Commun 5:3657.https://doi.org/10.1038/ncomms4657

Besnier F, Glover KA (2013) ParallelStructure: A R package to distribute parallel runs of the population genetics program STRUCTURE on multi-core computers. PLoS ONE 8(7): e70651. https://doi.org/10.1371/journal.pone.0070651

Bianco PG (2014) An update on the status of native and exotic freshwater fishes of Italy. J Appl Ichthyol 30: 62–77. https://doi.org/10.1111/jai.12291

Birnie-Gauvin K, Candee MM, Baktoft H, *et al* (2018) River connectivity reestablished: Effects and implications of six weir removals on brown trout smolt migration. River Res Appl 34:548–554. https://doi.org/10.1002/rra.3271

Blanc J, Chevassus B (1986) Survival growth and sexual maturation of the tiger trout hybrid (*Salmo trutta* ♀ ×*Salvelinus*

fontinalis ♂). Aquaculture 52:59–69. https://doi.org/10.1016/0044-8486(86)90108-0

Blanco-Lapaz À, Vergès JM (2016) Fish remains from the Neolithic site of El Mirador cave (Atapuerca, Spain): Seasonality and resource management. C R Palevol 15:745–751. https://doi.org/10.1016/j.crpv.2015.09.007

Blicharska M, Rönnbäck P (2018) Recreational fishing for sea trout—Resource for whom and to what value? Fish Res 204:380–389. https://doi.org/10.1016/j.fishres.2018.03.004

BOE (BOE-A-1907-7366) LEY del 15 de octubre de 1907 sobre pesca fluvial. Gaceta de Madrid No. 288, pp 195–196

BOE (BOE-A-1929-9983) REAL DECRETO-LEY del 24 de septiembre de 1929 introduciendo las variaciones que se indican en la legislación de pesca fluvial. Gaceta de Madrid No. 267, pp 1906–1913

BOE (BOE-A-1942-2205) LEY de 20 de febrero de 1942 por la que se regula el fomento y conservación de la pesca fluvial. Boletín Oficial del Estado No. 67, pp 1681 a 1694

BOE (BOE-A-1978-31229) CONSTITUCIÓN ESPAÑOLA, 29 de diciembre de 1978. No. 311, pp 29313–29424

BOE (BOE-A-2007-21490) LEY 42/2007, de 13 de diciembre, del Patrimonio Natural y de la Biodiversidad. Boletín Oficial del Estado No. 299, pp 51275–51327

Bohling J, Haffray P, Berrebi P (2016) Genetic diversity and population structure of domestic brown trout (*Salmo trutta*) in

France. Aquaculture 462:1–9. https://doi.org/10.1016/j.aquaculture.2016.04.013

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Boltana S, Sanhueza N, Donoso A, *et al* (2018) The expression of TRPV channels, prostaglandin E2 and pro-inflammatory cytokines during behavioural fever in fish. Brain Behav Immun 71:169–181. https://doi.org/10.1016/j.bbi.2018.03.023

Bouza C, Castro J, Sánchez L, Martínez P (1999) Genetic structure of brown trout, *Salmo trutta* L., at the southern limit of the distribution range of the anadromous form. Mol Ecol 8:1991–2001. https://doi.org/10.1046/j.1365-294x.1999.00794.x

Bouza C, Castro J, Sánchez L, Martínez P (2001) Allozymic evidence of parapatric differentiation of brown trout (*Salmo trutta* L.) within an Atlantic river basin of the Iberian Peninsula. Mol Ecol 10:1455–1469. https://doi.org/10.1046/j.1365-294X.2001.01272.x

Bouza C, Vilas R, Castro J, Martínez P (2008) Mitochondrial haplotype variability of brown trout populations from Northwestern Iberian Peninsula, a secondary contact area between lineages. Conserv Genet 9: 917–920. https://doi.org/10.1007/s10592-007-9398-8

Brown TA (2017) Genomes 4. 4th edn. Garland Science, New York.

Butler JRA, Radford A, Riddington G, Laughton R (2009) Evaluating an ecosystem service provided by Atlantic salmon, sea trout and other fish species in the River Spey, Scotland: The economic impact of recreational rod fisheries. Fish Res 96:259–266. https://doi.org/10.1016/j.fishres.2008.12.006

Caballero P, Morán P, Marco-Rius F (2013) A review of the genetic and ecological basis of phenotypic plasticity in brown trout. In: Polakov S, Moon, T (eds) Trout: From Physiology to Conservation. Nova Science, pp 9–26

Caballero P, Vieira-Lanero R, Cobo Fernando (2018) Sea trout (*Salmo trutta*) in Galicia (NW Spain). A Review. In: Lobón-Cerviá J and Sanz N (eds) Brown trout: Biology, ecology and management. Wiley, West Sussex, pp 445–482

Campbell MA, Buser TJ, Alfaro ME, López JA (2020) Addressing incomplete lineage sorting and paralogy in the inference of uncertain salmonid phylogenetic relationships. PeerJ 8:e9389. https://doi.org/10.7717/peerj.9389

Campos JL, Posada D, Morán P (2006) Genetic variation at MHC, mitochondrial and microsatellite loci in isolated populations of brown trout (*Salmo trutta*). Conserv Genet 7:515–530. https://doi.org/10.1007/s10592-005-9063-z

Carlsson J, Nilsson J (2001) Effects of geomorphological structures on genetic differentiation among brown trout populations in a northern boreal river drainage. Trans Am Fish Soc 130:36–45.

Caro T, Girling S (2010). Conservation by proxy. Island Press, Washington DC

Casacci LP, Barbero F, Balletto E (2014) The "Evolutionarily Significant Unit" concept and its applicability in biological conservation. Ital J Zool 81:182–193. https://doi.org/10.1080/11250003.2013.870240

Casal CM V. (2006) Global documentation of fish introductions: The growing crisis and recommendations for action. Biol Invasions 8:3–11. https://doi.org/10.1007/s10530-005-0231-3

Casillas S, Barbadilla A (2017) Molecular population genetics. Genetics 205:1003–1035. https://doi.org/10.1534/genetics.116.196493

Castro J, Pardo BG, Sánchez L, Martínez P (1999) rDNA RFLPs as genetic markers for resource management in brown trout. J Fish Biol 55:221–225. https://doi.org/10.1006/jfbi.1999.0982

Castro J, Pino A, Hermida M, *et al* (2006) A microsatellite marker tool for parentage analysis in Senegal sole (*Solea senegalensis*): Genotyping errors, null alleles and conformance to theoretical assumptions. Aquaculture 261:1194–1203. https://doi.org/10.1016/j.aquaculture.2006.09.001

Catchen JM, Amores A, Hohenlohe P, *et al* (2011) Stacks: Building and genotyping loci *de novo* from short-read sequences. G3-Genes Genom Genet 1(3):171–182. https://doi.org/10.1534/g3.111.000240

Catchen J, Hohenlohe PA, Bassham S, *et al* (2013) Stacks: An analysis tool set for population genomics. Mol Ecol 22:3124–3140. https://doi.org/10.1111/mec.12354

Christensen KA, Davidson WS (2017) Autopolyploidy genome duplication preserves other ancient genome duplications in Atlantic salmon (*Salmo salar*). PLoS ONE 12(2):e0173053. https://doi.org/10.1371/journal.pone.0173053

Cobo F, Couto MT, Gómez P, *et al* (2015). La trucha en la pesca a mosca. La biología de la trucha común. Sekotia, Madrid

Collins FS, Brooks LD, Chakravarti A (1999) A DNA polymorphism discovery resource for research on human genetic variation Genome Res 8:1229–1231

Comai L (2005) The advantages and disadvantages of being polyploid. Nat Rev Genet 6:836–846. https://doi.org/10.1038/nrg1711

Consejería de Agricultura, Medio Ambiente y Desarrollo Rural de Castilla la Mancha (2019) Plan de gestión de la trucha común en Castilla la Mancha https://www.castillalamancha.es/node/215379

Consellería de Medio Ambiente, Territorio e Vivenda de Galicia (2020) Serie histórica de licenzas de pesca fluvial 2000-2019 conforme á orixe do pescador. https://cmatv.xunta.gal/organizacion/c/CMAOT_DX_Conservacion_Natureza Accessed 28 May 2020

Cortey M, Pla C, García-Marín JL (2004) Historical biogeography of Mediterranean trout. Mol Phylogenet Evol 33(3):831–844. https://doi.org/10.1016/j.ympev.2004.08.012

Cortey M, Vera M, Pla C, García-Marín JL (2009) Northern and Southern expansions of Atlantic brown trout (*Salmo trutta*) populations during the Pleistocene. Biol J Linn Soc 97(4):904–917. https://doi.org/10.1111/j.1095-8312.2009.01220.x

Crête-Lafrenière A, Weir LK, Bernatchez L (2012) Framing the Salmonidae family phylogenetic portrait: A more complete picture from increased taxon sampling. PLoS ONE 7(10): e466627. https://doi.org/10.1371/journal.pone.0046662

Crutzen PJ, Stoermer EF (2000) The "Anthropocene". IGBP Newsletter 41:17–18

Cucherousset J, Aymes JC, Poulet N, *et al* (2008) Do native brown trout and non-native brook trout interact reproductively? Naturwissenschaften 95:647–654. https://doi.org/10.1007/s00114-008-0370-3

Curole JP, Hedgecock D (2007) Bivalve genomics: Complications, challenges, and future perspectives. In: Aquaculture Genome Technologies. Blackwell Publishing Ltd, Oxford, pp 525–543

Dana ED, García-de-Lomas J, González R, Ortega F (2011) Effectiveness of dam construction to contain the invasive crayfish *Procambarus clarkii* in a Mediterranean mountain stream. Ecol Eng 37:1607–1613.https://doi.org/10.1016/j.ecoleng.2011.06.014

Davey JW, Hohenlohe PA, Etter PD, *et al* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12:499–510. https://doi.org/doi:10.1038/nrg3012

Davidson WS, Huang TK, Fujiki K, *et al* (2009) The sex determining loci and sex chromosomes in the family Salmonidae. Sex Dev 3:78–87. https://doi.org/10.1159/000223073

De-Kayne R, Zoller S, Feulner PGD (2020) A *de novo* chromosome-level genome assembly of *Coregonus sp.* "Balchen": One representative of the Swiss Alpine whitefish radiation. Mol Ecol Resour 20:1093–1109. https://doi.org/10.1111/1755-0998.13187

Denic M, Geist J (2010) Habitat suitability analysis for lacustrine brown trout (*Salmo trutta*) in Lake Walchensee, Germany: Implications for the conservation of an endangered flagship species. Aquat Conserv Mar Freshw Ecosyst 20:9–17. https://doi.org/10.1002/aqc.1058

Diario Oficial del Ministerio de Marina (1912) LEY del 30 de octubre de 1912 sobre protección del salmón. No. 272, pp 1845–1849

Díaz-Arce N, Rodríguez-Ezpeleta N (2019) Selecting RAD-Seq data analysis parameters for population genetics: The more the better? Front Genet 10:533. https://doi.org/10.3389/fgene.2019.00533

Díaz-Rúa A, Chivite M, Comesaña S, *et al* (2020) The endocannabinoid system is affected by a high-fat-diet in rainbow

trout. Horm Behav 125:104825. https://doi.org/10.1016/j.yhbeh.2020.104825

Djurdjevič I, Kreft ME, Sušnik SB (2015) Comparison of pigment cell ultrastructure and organisation in the dermis of marble trout and brown trout, and first description of erythrophore ultrastructure in salmonids. J Anat 227:583–595. https://doi.org/10.1111/joa.12373

Do C, Waples RS, Peel D, *et al* (2014) NeEstimator v2: Re-implementation of software for the estimation of contemporary effective population size (N_e) from genetic data. Mol Ecol Resour 14:209–214. https://doi.org/10.1111/1755-0998.12157

Do C, Waples RS, Peel D, *et al* (2017) NeEstimator v. 2.1 Help file.

Doadrio I (2001) *Salmo trutta*. In: Doadrio I (ed) Atlas y libro rojo de los peces continentales de España, 2nd edn. Ministerio de Medio ambiente, Madrid, pp 129–131

Dodd JR, Cowx IG, Bolland JD (2017) Efficiency of a nature-like bypass channel for restoring longitudinal connectivity for a river-resident population of brown trout. J Environ Manage 204:318–326. https://doi.org/10.1016/j.jenvman.2017.09.004

DOG (1992) LEY 7/1992 del 24 de julio, de pesca fluvial de Galicia. Diario Oficial de Galicia. No., pp 6424–6433

DOG (2021) LEY 2/2021, de 8 de enero, de pesca continental de Galicia. Diario Oficial de Galicia. No. 9, pp 2244–2315

DOGC (2009) LEY 22/2009, de 23 de diciembre, de ordenación sostenible de la pesca en aguas continentales. Diari Oficial de la Generalitat de Catalunya. No. 5536, pp 97003–97031

Donoghue PCJ, Purnell MA (2005) Genome duplication, extinction and vertebrate evolution. Trends Ecol Evol 20(6):312–319. https://doi.org/10.1016/j.tree.2005.04.008

Dumas J, Bassenave JG, Jarry M, *et al* (2007) Effects of fish farm effluents on egg-to-fry development and survival of brown trout in artificial redds. J Fish Biol 70:1734–1758. https://doi.org/10.1111/j.1095-8649.2007.01442.x

Durrant CJ, Stevens JR, Hogstrand C, Bury NR (2011) The effect of metal pollution on the population genetic structure of brown trout (*Salmo trutta* L.) residing in the River Hayle, Cornwall, UK. Environ Pollut 159(12):3595–3603. https://doi.org/10.1016/j.envpol.2011.08.005

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour 4:359–361. https://doi.org/10.1007/s12686-011-9548-7

Eaton DAR, Overcast I (2020) ipyrad: Interactive assembly and analysis of RADseq datasets. Bioinformatics 36(8):2592–2594. https://doi.org/10.1093/bioinformatics/btz966

Eddy FB, Handy RD (2013) Ecological and Environmental Physiology of Fishes. Oxford University Press, Oxford

Eklöv AG, Greenberg LA, Brönmark C, *et al* (1999) Influence of water quality, habitat and species richness on brown trout populations. J Fish Biol 54:33–43. https://doi.org/10.1006/jfbi.1998.0842

Ellstrand NC, Elam DR (1993) Population genetic consequences of small population size: Implications for plant conservation. Annu Rev Ecol Syst 24:217–242. https://doi.org/10.1146/annurev.es.24.110193.001245

England PR, Cornuet JM, Berthier P, *et al* (2006) Estimating effective population size from linkage disequilibrium: Severe bias in small samples. Conserv Genet 7:303–308. https://doi.org/10.1007/s10592-005-9103-8

Eslamloo K, Xue X, Hall JR, *et al* (2017) Transcriptome profiling of antiviral immune and dietary fatty acid dependent responses of Atlantic salmon macrophage-like cells. BMC Genomics 18:706. https://doi.org/10.1186/s12864-017-4099-2

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. Mol Ecol 14:2611–2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics. 131(2):479–491. https://doi.org/10.1093/genetics/131.2.479

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. Heredity. 103:285–298. https://doi.org/10.1038/hdy.2009.74

Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour. 10(3):564–567. https://doi.org/10.1111/j.1755-0998.2010.02847.x

Excoffier L, Laval G (2015) Arlequin v. 3.5.2 user manual: an integrated software package for population genetics data analysis. Bernem, Switzerland: University of Berne

Fan G, Song Y, Yang L, *et al* (2020) Initial data release and announcement of the 10,000 Fish Genomes Project (Fish10K). Gigascience 9(8):giaa080. https://doi.org/10.1093/gigascience/giaa080

Ferguson A (1989) Genetic differences among brown trout, *Salmo trutta*, stocks and their importance for the conservation and management of the species. Freshw Biol 21:35–46. https://doi.org/10.1111/j.1365-2427.1989.tb01346.x

Ferguson A, Reed TE, Cross TF, *et al* (2019) Anadromy, potamodromy and residency in brown trout *Salmo trutta*: the role of genes and the environment. J Fish Biol 95:692–718. https://doi.org/10.1111/jfb.14005

Fernández-Cebrián R, Araguas RM, Sanz N, García-Marín JL (2014) Genetic risks of supplementing trout populations with native stocks: A simulation case study from current Pyrenean

populations. Can J Fish Aquat Sci 71:1243–1255. https://doi.org/10.1139/cjfas-2013-0362

Fitzpatrick BM (2009) Power and sample size for nested analysis of molecular variance. Mol Ecol 18:3961–3966. https://doi.org/10.1111/j.1365-294X.2009.04314.x

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. Genetics 180(2):977–993. https://doi.org/10.1534/genetics.108.092221

Food and Agriculture Organization of the United Nations (2012) FAO technical guidelines for responsible fisheries 13 Recreational fisheries. Rome

Fountain ED, Pauli JN, Reid BN, *et al* (2016) Finding the right coverage: the impact of coverage and sequence quality on single nucleotide polymorphism genotyping error rates. Mol Ecol Resour 16:966–978. https://doi.org/10.1111/1755-0998.12519

Frankham R, Ballou J, Briscoe, D (2010). Introduction to conservation genetics. 2nd edn. Cambridge University Press, New York

Frankham R, Bradshaw CJA, Brook BW (2014) Genetics in conservation management: Revised recommendations for the 50/500 rules, Red List criteria and population viability analyses. Biol Conserv 170:56–63. https://doi.org/10.1016/j.biocon.2013.12.036

Franklin, I.R., 1980. Evolutionary change in small populations. In: Soulé, M.E.,Wilcox, B.A. (Eds.), Conservation

Biology: An Evolutionary-Ecological Perspective. Sinauer, Sunderland, MA, pp. 135–149

Freeland J, Kirk H, Petersen S (2011). Molecular Ecology. 2nd edn. Wiley-Blackwell, West Sussex

Freyhof, J (2011) *Salmo trutta*. The IUCN Red List of Threatened Species 2011: e.T19861A9050312. https://dx.doi.org/10.2305/IUCN.UK.2008.RLTS.T19861A9050312.en. Accessed 12 January 2020

Fricke R, Eschmeyer W, Fong JD. CAS - Eschmeyer's Catalog of Fishes - Species by Family. 2020.http://researcharchive.calacademy.org/research/ichthyology/catalog/SpeciesByFamily.asp. Accessed 14 Jul 2020

Froehlich HE, Gentry RR, Halpern BS (2017) Conservation aquaculture: Shifting the narrative and paradigm of aquaculture's role in resource management. Biol Conserv 215:162–168. https://doi.org/10.1016/j.biocon.2017.09.012

Fu L, Niu B, Zhu Z, *et al* (2012) CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics 28(23):3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Gaggiotti OE, Chao A, Peres-Neto P, *et al* (2018) Diversity from genes to ecosystems: A unifying framework to study variation across biological metrics and scales. Evol Appl 11:1176–1193. https://doi.org/10.1111/eva.12593

Galla SJ, Forsdick NJ, Brown L, *et al* (2019) Reference genomes from distantly related species can be used for discovery

of single nucleotide polymorphisms to inform conservation management. Genes 10:9. https://doi.org/10.3390/genes10010009

Garcia-Marin JL, Jorde PE, Ryman N, *et al* (1991) Management implications of genetic differentiation between native and hatchery populations of brown trout (*Salmo trutta*) in Spain. Aquaculture 95(3-4):235–249. https://doi.org/10.1016/0044-8486(91)90090-T

García-Marín JL, Pla C (1996) Origins and relationships of native populations of *Salmo trutta* (brown trout) in Spain. Heredity 77:313–323. https://doi.org/10.1038/hdy.1996.145

García-Marín JL, Araguas RM, Vera M, *et al* (2018) Understanding the brown trout population genetic structure in the Iberian Peninsula. A Review. In: Lobón-Cerviá J and Sanz N (eds) Brown trout: Biology, ecology and management. Wiley, West Sussex, pp 103–126

Geza E, Mugo J, Mulder NJ, *et al* (2019) A comprehensive survey of models for dissecting local ancestry deconvolution in human genome. Brief Bioinform 20(5):1709–1724. https://doi.org/10.1093/bib/bby044

Gilbert KJ, Andrew RL, Bock DG, *et al* (2012) Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program STRUCTURE. Mol Ecol, 21: 4925–4930. https://doi.org/10.1111/j.1365-294X.2012.05754.x

Glaubitz JC, Casstevens TM, Lu F, *et al* (2014) TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline.

PLoS ONE 9 (2): e90346. https://doi.org/10.1371/journal.pone.0090346

Glover NM, Redestig H, Dessimoz C (2016) Homoeologs: What are they and how do we infer them? Trends Plant Sci 21(7):609–621. https://doi.org/10.1016/j.tplants.2016.02.005

Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17:333–351. https://doi.org/10.1038/nrg.2016.49

Götz S, García-Gómez JM, Terol J, *et al* (2008) High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 36(10):3420–3435. https://doi.org/10.1093/nar/gkn176

Gratton P, Allegrucci G, Sbordoni V, Gandolfi A (2014) The evolutionary jigsaw puzzle of the surviving trout (*Salmo trutta* L. complex) diversity in the Italian region. A multilocus bayesian approach. Mol Phylogenet Evol 79:292–304. https://doi.org/10.1016/j.ympev.2014.06.022

Grimholt U, Lukacs M (2021) MHC class I evolution; from Northern pike to salmonids. BMC Ecol Evo 21:3. https://doi.org/10.21203/rs.3.rs-36819/v1

Guillot G, Estoup A, Mortier F, Cosson JF (2005) A spatial statistical model for landscape genetics. Genetics 170(3):1261–1280. https://doi.org/10.1534/genetics.104.033803

Hall, TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids. Symp. Ser. 41:95–98

Hamilton KE, Ferguson A, Taggart JB (1989) Post-glacial colonization of brown trout, *Salmo trutta* L.: *Ldh-5* as a phylogeographic marker locus. J Fish Biol 35:651–664. https://doi.org/10.1111/j.1095-8649.1989.tb03017.x

Hansen MM, Nielsen EE, Bekkevold D, Mensberg K-LD (2001) Admixture analysis and stocking impact assessment in brown trout (*Salmo trutta*), estimated with incomplete baseline data. Can J Fish Aquat Sci 58:1853–1860. https://doi.org/10.1139/cjfas-58-9-1853

Hansen T, Fjelldal PG, Lien S *et al* (2021) The genome sequence of the brown trout, *Salmo trutta* Linnaeus 1758 [version 1; peer review: 1 approved]. Wellcome Open Res 6:108. https://doi.org/10.12688/wellcomeopenres.16838.1

Hao Z, Lv D, Ge Y, *et al* (2020) *RIdeogram*: Drawing SVG graphics to visualize and map genome-wide data on the idiograms. PeerJ Comput Sci 6:e251. https://doi.org/10.7717/peerj-cs.251

Hari RE, Livingstone DM, Siber R, *et al* (2006) Consequences of climatic change for water temperature and brown trout populations in Alpine rivers and streams. Glob Chang Biol 12:10–26. https://doi.org/10.1111/j.1365-2486.2005.001051.x

Hays CG, Fagan CM (2016) Evolution and Conservation Biology In: Kliman R (ed) Encyclopedia of evolutionary biology. Elsevier, Oxford, pp 347–353

Heggenes J, Røed KH (2006) Do dams increase genetic diversity in brown trout (*Salmo trutta*)? Microgeographic

differentiation in a fragmented river. Ecol Freshw Fish. 15: 366–375. https://doi.org/10.1111/j.1600-0633.2006.00146.x

Hendricks S, Anderson EC, Antao T, *et al* (2018) Recent advances in conservation and population genomics data analysis. Evol Appl 11:1197–1211. https://doi.org/10.1111/eva.12659

Hermida M, San Miguel E, Bouza C, *et al* (2009) Morphological variation in a secondary contact between divergent lineages of brown trout (*Salmo trutta*) from the Iberian Peninsula. Genet Mol Biol 32(1):42–50. https://doi.org/10.1590/S1415-47572009005000014

Hill RW, Wyse GA, Anderson M (2012) Animal Physiology, 4th edn. Sinauer Associates, Sunderland

Hirsch CD, Evans J, Buell CR, Hirsch CN (2014) Reduced representation approaches to interrogate genome diversity in large repetitive plant genomes. Briefings Funct Genomics Proteomics 13(4):257–267. https://doi.org/10.1093/bfgp/elt051

Ho A, Murphy M, Wilson S, *et al* (2011) Sequencing by ligation variation with endonuclease V digestion and deoxyinosine-containing query oligonucleotides. BMC Genomics 12:598. https://doi.org/10.1186/1471-2164-12-598

Hodel RGJ, Chen S, Payton AC, *et al* (2017) Adding loci improves phylogeographic resolution in red mangroves despite increased missing data: Comparing microsatellites and RAD-Seq and investigating loci filtering. Sci Rep 7:17598. https://doi.org/10.1038/s41598-017-16810-7

Huang YF, Chen SC, Chiang YS, *et al* (2012) Palindromic sequence impedes sequencing-by-ligation mechanism. BMC Syst Biol 6:S10. https://doi.org/10.1186/1752-0509-6-S2-S10

Hubbs CL (1955) Hybridization between fish species in nature. Syst Zool 4:1–20 https://doi.org/10.2307/sysbio/4.1.1

Huber M (2010) Compendium of Bivalves. A Full-color Guide to 3,300 of the World's Marine Bivalves. A Status on Bivalvia after 250 Years of Research. ConchBooks, Hackenheim

Hughes AR, Stachowicz JJ (2004) Genetic diversity enhances the resistance of a seagrass ecosystem to disturbance. Proc Natl Acad Sci USA 101(24):8998–9002. https://doi.org/10.1073/pnas.0402642101

Hughes AR, Inouye BD, Johnson MTJ, *et al* (2008) Ecological consequences of genetic diversity. Ecol Lett 11:609–623. https://doi.org/10.1111/j.1461-0248.2008.01179.x

Inland Fisheries Ireland (2013) Socio-Economic Study of Recreational Angling in Ireland. Tourism development international, Dún Laoghaire

Jacobs A, Hughes MR, Robinson PC, *et al* (2018) The genetic architecture underlying the evolution of a rare piscivorous life history form in brown trout after secondary contact and strong introgression. Genes 9(6):280 https://doi.org/10.3390/genes9060280

Jamalzadeh HR, Akhundian M, Kabir M, *et al* (2013) Changes of gonadosomatic index and plasma levels of cortisol in the male and female captive Caspian brown trout, *Salmo Trutta*

Caspius (Kessler, 1877) during the reproductive cycle. J Appl Anim Res 41(2):133–136. https://doi.org/10.1080/09712119.2012.738228

Jensen LF, Hansen MM, Pertoldi C, *et al* (2008) Local adaptation in brown trout early life-history traits: Implications for climate change adaptability. Proc R Soc B Biol Sci. 275:2859–2868. https://doi.org/10.1098/rspb.2008.0870

Jiang D, Chen J, Fan Z, *et al* (2017) CRISPR/Cas9-induced disruption of *wt1a* and *wt1b* reveals their different roles in kidney and gonad development in Nile tilapia. Dev Biol 428:63–73. https://doi.org/10.1016/j.ydbio.2017.05.017

Jiggins CD, Mallet J (2000) Bimodal hybrid zones and speciation. Trends Ecol Evol 15:250–255. https://doi.org/10.1016/S0169-5347(00)01873-5

Johnstone R, Simpson TH, Youngson AF (1978) Sex reversal in salmonid culture. Aquaculture13(2):115–134. https://doi.org/10.1016/0044-8486(78)90106-0

Jombart T (2008) *adegenet*: A R package for the multivariate analysis of genetic markers. Bioinformatics 24(11):1403–1405 https://doi.org/10.1093/bioinformatics/btn129

Jombart T, Ahmed I (2011) *adegenet 1.3-1*: New tools for the analysis of genome-wide SNP data. Bioinformatics 27(21):3070–3071 https://doi.org/10.1093/bioinformatics/btr521

Jones AT, Ovenden JR, Wang YG (2016) Improved confidence intervals for the linkage disequilibrium method for

estimating effective population size. Heredity 117:217–223. https://doi.org/10.1038/hdy.2016.19

Jonsson B, Jonsson N (2011) Ecology of Atlantic salmon and brown trout. Habitat as a template for life histories. Springer, London

Jorde PE, Ryman N (2007) Unbiased estimator for genetic drift and effective population size. Genetics 177(2):927–935. https://doi.org/10.1534/genetics.107.075481

Kalayci G, Ozturk RC, Capkin E, Altinok I (2018) Genetic and molecular evidence that brown trout *Salmo trutta* belonging to the Danubian lineage are a single biological species. J Fish Biol 93:792–804. https://doi.org/10.1111/jfb.13777

Karakousis Y, Triantaphyllidis C, Economidis PS (1991) Morphological variability among seven populations of brown trout, *Salmo trutta* L., in Greece. J Fish Biol 38:807–817. https://doi.org/10.1111/j.1095-8649.1991.tb03620.x

Keenan K, Mcginnity P, Cross TF, *et al* (2013) DiveRsity: An R package for the estimation and exploration of population genetics parameters and their associated errors. Methods Ecol Evol 4:782–788. https://doi.org/10.1111/2041-210X.12067

Klemetsen A, Amundsen PA, Dempson JB, *et al* (2003) Atlantic salmon *Salmo salar* L., brown trout *Salmo trutta* L. and Arctic charr *Salvelinus alpinus* (L.): A review of aspects of their life histories. Ecol Freshw Fish 12:1–59. https://doi.org/10.1034/j.1600-0633.2003.00010.x

Kopelman NM, Mayzel J, Jakobsson M, *et al* (2015) Clumpak: A program for identifying clustering modes and packaging population structure inferences across *K*. Mol Ecol Resour 15:1179–1191. https://doi.org/10.1111/1755-0998.12387

Kottelat M, Freyhof J (2007) Family Salmonidae. In: Kottelat M, Freyhof J (eds) Handbook of European freshwater fishes. Publications Kottelat, Berlin, pp 395–457

Lamas S, Fernández JA, Aboal JR, Carballeira A (2007) Testing the use of juvenile *Salmo trutta* L. as biomonitors of heavy metal pollution in freshwater. Chemosphere 67(2):221–228. https://doi.org/10.1016/j.chemosphere.2006.10.040

Langeland A, Nøst T (1995) Gill raker structure and selective predation on zooplankton by particulate feeding fish. J Fish Biol 47:719–732. https://doi.org/10.1111/j.1095-8649.1995.tb01937.x

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25 https://doi.org/10.1186/gb-2009-10-3-r25

Lankau RA, Strauss SY (2007) Mutual feedbacks maintain both genetic and species diversity in a plant community. Science 317:1561–1563. https://doi.org/10.1126/science.1147455

Leaniz CG, Verspoor E (1989) Natural hybridization between Atlantic salmon, *Salmo salar*, and brown trout, *Salmo trutta*, in northern Spain. J Fish Biol 34:41-46. https://doi.org/10.1111/j.1095-8649.1989.tb02956.x

Lecaudey LA, Schliewen UK, Osinov AG, *et al* (2018) Inferring phylogenetic structure, hybridization and divergence times within Salmoninae (Teleostei: Salmonidae) using RAD-sequencing. Mol Phylogenet Evol 124:82–99. https://doi.org/10.1016/j.ympev.2018.02.022

Leclercq E, Taylor JF, Migaud H (2010) Morphological skin colour changes in teleosts. Fish and Fisheries 11:159–193. https://doi.org/10.1111/j.1467-2979.2009.00346.x

Lehnert SJ, Baillie SM, MacMillan J, *et al* (2020) Multiple decades of stocking has resulted in limited hatchery introgression in wild brook trout (*Salvelinus fontinalis*) populations of Nova Scotia. Evol Appl 13:1069–1089. https://doi.org/10.1111/eva.12923

Leitwein M, Guinand B, Pouzadoux J, *et al* (2017) A dense brown trout (*Salmo trutta*) linkage map reveals recent chromosomal rearrangements in the Salmo genus and the impact of selection on linked neutral diversity. G3-Genes Genom Genet 7(4):1365–1376. https://doi.org/10.1534/g3.116.038497

Leitwein M, Gagnaire PA, Desmarais E, *et al* (2018) Genomic consequences of a recent three-way admixture in supplemented wild brown trout populations revealed by local ancestry tracts. Mol Ecol 27:3466–3483. https://doi.org/10.1111/mec.14816

Lemopoulos A, Uusi-Heikkilä S, Huusko A, *et al* (2018) Comparison of migratory and resident populations of brown trout reveals candidate genes for migration tendency. Genome Biol Evol 10(6):1493–1503. https://doi.org/10.1093/gbe/evy102

Lemopoulos A, Prokkola JM, Uusi-Heikkilä S, *et al* (2019a) Comparing RAD-Seq and microsatellites for estimating genetic diversity and relatedness — Implications for brown trout conservation. Ecol Evol 9:2106–2120. https://doi.org/10.1002/ece3.4905

Lemopoulos A, Uusi-Heikkilä S, Hyvärinen P, *et al* (2019b) Association mapping based on a common-garden migration experiment reveals candidate genes for migration tendency in brown trout. G3-Genes Genom Genet 9(9):2887–2896. https://doi.org/10.1534/g3.119.400369

Lessard JL, Hayes DB (2003) Effects of elevated water temperature on fish and macroinvertebrate communities below small dams. River Res Appl. 19: 721–732. https://doi.org/10.1002/rra.713

Lewin HA, Robinson GE, Kress WJ, *et al* (2018) Earth BioGenome Project: Sequencing life for the future of life. Proc Natl Acad Sci 115(17):4325–4333. https://doi.org/10.1073/pnas.1720115115

Li H, Handsaker B, Wysoker A, Fennell T, *et al* The Sequence alignment/map format and SAMtools, Bioinformatics (2009) 25(16) 2078-2079 https://doi.org/10.1093/bioinformatics/btp352

Li W, Godzik A (2006) Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13):1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Li YL, Liu JX (2018) StructureSelector: A web-based software to select and visualize the optimal number of clusters using multiple methods. Mol Ecol Resour 18:176–177. https://doi.org/10.1111/1755-0998.12719

Linck E, Battey CJ (2019) Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. Mol Ecol Resour 19:639–647. https://doi.org/10.1111/1755-0998.12995

Linde AR, Sánchez-Galán S, Izquierdo JI, *et al* (1998) Brown trout as biomonitor of heavy metal pollution: Effect of age on the reliability of the assessment. Ecotoxicology and Environmental Safety 40(1-2), 120–125 https://doi.org/10.1006/eesa.1998.1652

Linde AR, Sanchez-Galan S, Garcia-Vazquez E (2004) Heavy metal contamination of European eel (*Anguilla anguilla*) and brown trout (*Salmo trutta*) caught in wild ecosystems in Spain. J Food Prot. 67(10):2332–2336. https://doi.org/10.4315/0362-028X-67.10.2332

Linløkken AN, Haugen TO, Kent MP, Lien S (2017) Genetic differences between wild and hatchery-bred brown trout (*Salmo trutta* L.) in single nucleotide polymorphisms linked to selective traits. Ecol Evol 7:4963–4972. https://doi.org/10.1002/ece3.3070

Lischer HE, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics 28(2):298–299. https://doi.org/10.1093/bioinformatics/btr642

Lobón-Cerviá, J (2018) Princess of the streams: The brown trout *Salmo trutta* L. as aquatic royalty. In: Lobón-Cerviá J and Sanz N (eds) Brown trout: Biology, ecology and management. Wiley, West Sussex, pp 1–14

Lopes-Lima M, Sousa R, Geist J, *et al* (2017) Conservation status of freshwater mussels in Europe: state of the art and future challenges. Biol Rev 92:572–607. https://doi.org/10.1111/brv.12244

Louis EJ, Dempster ER (1987) An Exact Test for Hardy-Weinberg and Multiple Alleles. Biometrics 43(4):805–811

Luca F, Hudson RR, Witonsky DB, Di Rienzo A (2011) A reduced representation approach to population genetic analyses and applications to human evolution. Genome Res 21:1087–1098. https://doi.org/10.1101/gr.119792.110

Luckenbach T, Triebskorn R, Müller E, Oberemm A (2001) Toxicity of waters from two streams to early life stages of brown trout (*Salmo trutta* f. fario L.), tested under semi-field conditions. Chemosphere 45(4-5):571-579. https://doi.org/10.1016/S0045-6535(00)00595-6

Lundberg P (1988) The evolution of partial migration in birds. Trends Ecol Evol 3(7):172–175. https://doi.org/10.1016/0169-5347(88)90035-3

Luque GM, Vayssade C, Facon B, *et al* (2016) The genetic Allee effect: A unified framework for the genetics and demography of small populations. Ecosphere 7(7):e01413 https://doi.org/10.1002/ecs2.1413

Machordom A, Suárez J, Almodóvar A, *et al* (2000) Mitochondrial haplotype variation and phylogeography of Iberian brown trout populations. Mol Ecol 9:1324–1338. https://doi.org/10.1046/j.1365-294x.2000.01015.x

Macqueen DJ, Johnston IA (2014) A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. Proc R Soc B Biol Sci 281:2881. https://doi.org/10.1098/rspb.2013.2881

Mantere T, Kersten S, Hoischen A (2019) Long-read sequencing emerging in medical genetics. Front Genet 10:426. https://doi.org/10.3389/fgene.2019.00426

Manuzzi A, Zane L, Muñoz-Merida A, *et al* (2019) Population genomics and phylogeography of a benthic coastal shark (*Scyliorhinus canicula*) using 2b-RAD single nucleotide polymorphisms. Biol J Linn Soc 126(2):289–303. https://doi.org/10.1093/biolinnean/bly185

Marandel F, Lorance P, Berthelé O, *et al* (2019) Estimating effective population size of large marine populations, is it feasible? Fish Fish 20:189–198. https://doi.org/10.1111/faf.12338

Marandel F, Charrier G, Lamy JB, *et al* (2020) Estimating effective population size using RADseq: Effects of SNP selection and sample size. Ecol Evol 10:1929–1937. https://doi.org/10.1002/ece3.6016

Marco-Rius F, Sotelo G, Caballero P, Morán P (2013) Insights for planning an effective stocking program in anadromous

brown trout (*Salmo trutta*). Can J Fish Aquat Sci 70:7. https://doi.org/10.1139/cjfas-2013-0084

Maroso F, Casanova A, do Prado FD, *et al* (2018) Species identification of two closely exploited flatfish, turbot (*Scophthalmus maximus*) and brill (*Scophthalmus rhombus*), using a ddRAD-Seq genomic approach. Aquat Conserv Mar Freshw Ecosyst 28:1253–1260. https://doi.org/10.1002/aqc.2932

Maroso F, De Gracia CP, Iglesias D, *et al* (2019) A useful SNP panel to distinguish two cockle species, *Cerastoderma edule* and *C. glaucum*, co-occurring in some european beds, and their putative hybrids. Genes.10(10):760. https://doi.org/10.3390/genes10100760

Martínez P, Arias J, Castro J, Sánchez L (1993) Differential stocking incidence in brown trout (*Salmo trutta*) populations from Northwestern Spain. Aquaculture 114(3–4):203–216. https://doi.org/10.1016/0044-8486(93)90296-B

Martínez P, Bouza C, Castro J, *et al* (2007) Analysis of a secondary contact between divergent lineages of brown trout *Salmo trutta* L. from Duero basin using microsatellites and mtDNA RFLPs. J Fish Biol 71:195–213. https://doi.org/10.1111/j.1095-8649.2007.01551.x

Martínez P, Castro J, Pardo BG, et al (2009) High Ag-NOR-site variation associated to a secondary contact in brown trout from the Iberian Peninsula. Genetica 136:419–427. https://doi.org/10.1007/s10709-008-9342-y

Martínez L, Freire R, Arias - Pérez A, *et al* (2015) Patterns of genetic variation across the distribution range of the cockle *Cerastoderma edule* inferred from microsatellites and mitochondrial DNA. Mar Biol 162:1393–1406. https://doi.org/10.1007/s00227-015-2676-y

Mastretta-Yanes A, Arrigo N, Álvarez N, *et al* (2015) Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference. Mol Ecol Resour 15:28–41. https://doi.org/10.1111/1755-0998.12291

Meirmans PG (2015) Seven common mistakes in population genetics and how to avoid them. Mol Ecol 24:3223–3231. https://doi.org/10.1111/mec.13243

Mignardi M, Nilsson M (2014) Fourth-generation sequencing in the cell and the clinic. Genome Med 6:31. https://doi.org/10.1186/gm548

Milan M, Maroso F, Dalla Rovere G, *et al* (2019) Tracing seafood at high spatial resolution using NGS-generated data and machine learning: Comparing microbiome versus SNPs. Food Chem 286:413–420. https://doi.org/10.1016/j.foodchem.2019.02.037

Millennium Ecosystem Assessment (2005) Ecosystems and Human Well-being: Synthesis. Island Press, Washington, DC

Ministerio para la Transición Ecológica y el Reto Demográfico, MITERD, (2020) Anuario de estadística forestal 2018. MITERD, Madrid.

Moorkens E, Cordeiro J, Seddon MB, *et al* (2017) *Margaritifera margaritifera*. The IUCN Red List of Threatened Species 2017: e.T12799A128686456. https://dx.doi.org/10.2305/IUCN.UK.2017-3.RLTS.T12799A508865.en . Accessed 12 January 2020

Morán P, Pendás AM, Garcia-Vázquez E, Izquierdo J (1991) Failure of a stocking policy, of hatchery reared brown trout, *Salmo trutta* L., in Asturias, Spain, detected using *LDH-5** as a genetic marker. J Fish Biol 39:117–121 https://doi.org/10.1111/j.1095-8649.1991.tb05075.x

Morán P, Marco-Rius F, Megías M, *et al* (2013) Environmental induced methylation changes associated with seawater adaptation in brown trout. Aquaculture 392–395:77–83. https://doi.org/10.1016/j.aquaculture.2013.02.006

Moritz C (1994) Defining "Evolutionarily Significant Units" for conservation. Trends Ecol Evol 9:373–375. https://doi.org/10.1016/0169-5347(94)90057-4

Nakatani Y, Takeda H, Kohara Y, Morishita S (2011) Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. In: Naruse K, Tanaka M, Takeda H (eds) Medaka a model for organogenesis, human disease, and evolution. Springer, Tokyo pp 307–322

Narum SR, Hess JE (2011) Comparison of F_{ST} outlier tests for SNP loci under selection. Mol Ecol Resour 11:184–194. https://doi.org/10.1111/j.1755-0998.2011.02987.x

NCBI *Salmo trutta* assembly: https://www.ncbi.nlm.nih.gov/assembly/GCF_901001165.1. Accessed 26 July 2020

Nei M, Tajima F (1981) Genetic drift and estimation of effective population size. Genetics 98(3):625–640. https://doi.org/10.1093/genetics/98.3.625

Nugent CM, Easton AA, Norman JD, *et al* (2017) A SNP based linkage map of the arctic charr (*Salvelinus alpinus*) genome provides insights into the diploidization process after whole genome duplication. G3-Genes Genom Genet 7(2):543–556. https://doi.org/10.1534/g3.116.038026

Nusbaumer D, Marques da Cunha L, Wedekind C (2020) Testing for population differences in evolutionary responses to pesticide pollution in brown trout (*Salmo trutta*). Evol Appl 14:462– 475. https://doi.org/10.1111/eva.13132

O'Connor JD, Murphy S, Lally HT, *et al* (2020) Microplastics in brown trout (*Salmo trutta* Linnaeus, 1758) from an Irish riverine system. Environ Pollut 267:115572. https://doi.org/10.1016/j.envpol.2020.115572

O'Leary SJ, Puritz JB, Willis SC, *et al* (2018) These aren't the loci you're looking for: Principles of effective SNP filtering for molecular ecologists. Mol Ecol 27:3193–3206. https://doi.org/10.1111/mec.14792

OESA - Fundación Biodiversidad (2018) Caracterización de la cría en cautividad y repoblación de especies de interés a través de la acuicultura. Fundación Biodiversidad, Madrid

Ohno, D.S. (1970). Evolution by Gene Duplication. Springer, New York

Ojanguren AF, Braña F (2003) Thermal dependence of embryonic growth and development in brown trout. J Fish Biol 62:580–590. https://doi.org/10.1046/j.1095-8649.2003.00049.x

OJEC (1978) Council Directive 78/659/EEC of 18 July 1978 on the quality of fresh waters needing protection or improvement in order to support fish life. Official Journal of the European Community No. L222, pp 1–10

OJEC (2000) Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy Official Journal of the European Community No. L327, pp 1–73

Östergren J, Nilsson J (2012) Importance of life-history and landscape characteristics for genetic structure and genetic diversity of brown trout (*Salmo trutta* L.). Ecol Freshw Fish 21:119–133. https://doi.org/10.1111/j.1600-0633.2011.00529.x

Ősz, Á, Horváth, Á, Hoitsy, G, *et al* (2018) The genetic status of the Hungarian brown trout populations: exploration of a blind spot on the European map of *Salmo trutta* studies. PeerJ 6:e5152. https://doi.org/10.7717/peerj.5152

Paetkau D, Slade R, Burden M, *et al* (2004) Genetic assignment methods for the direct, real-time estimation of migration rate: A simulation-based exploration of accuracy and power. Mol Ecol 13:55–65. https://doi.org/10.1046/j.1365-294X.2004.02008.x

Paladini G, Shinn AP, Taylor NGH, *et al* (2021) Geographical distribution of *Gyrodactylus salaris* Malmberg, 1957 (Monogenea, Gyrodactylidae). Parasite Vector 14:34. https://doi.org/10.1186/s13071-020-04504-5

Palm S, Laikre L, Jorde PE, Ryman N (2003) Effective population size and temporal genetic change in stream resident brown trout (*Salmo trutta*, L.). Conserv Genet 4:249–264. https://doi.org/10.1023/A:1024064913094

Paris JR, Stevens JR, Catchen JM (2017) Lost in parameter space: a road map for STACKS. Methods Ecol Evol 8:1360–1373. https://doi.org/10.1111/2041-210X.12775

Pasquier J, Cabau C, Nguyen T, *et al* (2016) Gene evolution and gene expression after whole genome duplication in fish: The PhyloFish database. BMC Genomics 17: 368. https://doi.org/10.1186/s12864-016-2709-z

Pembleton LW, Cogan NOI, Forster JW (2013) StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. Mol Ecol Resour 13:946–952. https://doi.org/10.1111/1755-0998.12129

Perdices A, Bermingham E, Montilla A, Doadrio I (2002) Evolutionary history of the genus *Rhamdia* (Teleostei: Pimelodidae) in Central America. Mol Phylogenet Evol 25:172–189. https://doi.org/10.1016/S1055-7903(02)00224-5

Pickering AD (1989) Environmental stress and the survival of brown trout, *Salmo trutta*. Freshw Biol 21:47–55. https://doi.org/10.1111/j.1365-2427.1989.tb01347.x

Pinter K, Epifanio J, Unfer G (2019) Release of hatchery-reared brown trout (*Salmo trutta*) as a threat to wild populations? A case study from Austria. Fish Res 219:105296. https://doi.org/10.1016/j.fishres.2019.05.013

Piry S, Alapetite A, Cornuet JM, *et al* (2004) GENECLASS2: A software for genetic assignment and first-generation migrant detection. J Hered 95(6):536–539. https://doi.org/10.1093/jhered/esh074

Pollak E (1983) A new method for estimating the effective population size from allele frequency changes. Genetics 104(3):531–548. https://doi.org/10.1093/genetics/104.3.531

Porras-Hurtado L, Ruiz Y, Santos C, *et al* (2013) An overview of *STRUCTURE*: Applications, parameter settings, and supporting software. Front Genet 4:98. https://doi.org/10.3389/fgene.2013.00098

Pouzadoux J, Guinand B, Desmarais E, *et al* (2017) A dense brown trout (*Salmo trutta*) linkage map reveals recent chromosomal rearrangements in the *Salmo* genus and the impact of selection on linked neutral diversity. G3-Genes Genom Genet 7(4):1365–1376. https://doi.org/10.1534/g3.116.038497

Prado FD, Vera M, Hermida M, *et al* (2018) Tracing the genetic impact of farmed turbot *Scophthalmus maximus* on wild populations. Aquac Environ Interact 10:447–463. https://doi.org/10.3354/AEI00282

Presa P, Pardo BG, Martínez P, Bernatchez L (2002) Phylogeographic congruence between mtDNA and rDNA ITS

markers in brown trout. Mol Biol Evol 19(12):2161–2175. https://doi.org/10.1093/oxfordjournals.molbev.a004041

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–59

Puechmaille SJ (2016) The program structure does not reliably recover the correct population structure when sampling is uneven: Subsampling and new estimators alleviate the problem. Mol Ecol Resour 16:608–627. https://doi.org/10.1111/1755-0998.12512

Puritz JB, Hollenbeck CM, Gold JR (2014) dDocent: A RAD-Seq, variant-calling pipeline designed for population genomics of non-model organisms. PeerJ 2:e431. https://doi.org/10.7717/peerj.431

Pustovrh G, Snoj A, Bajec SS (2011) A set of SNPs enabling identification of trouts and their hybrids in *Salmo* genus. Conserv Genet Resour 3:147–150. https://doi.org/10.1007/s12686-010-9310-6

Quail MA, Smith M, Coupland P, *et al* (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13:341. https://doi.org/10.1186/1471-2164-13-341

Reyes-Gavilán FG, Ojanguren AF, Brana F (1997) The ontogenetic development of body segments and sexual dimorphism in brown trout (*Salmo trutta* L.). Can J Zool 75:651–655. https://doi.org/10.1139/z97-083

Ríos N, Casanova A, Hermida M, *et al* (2020) Population Genomics in *Rhamdia quelen*. Divergence and adaptation in the Neotropical region. Genes 11: 109. https://doi.org/10.3390/genes11010109

Ryder OA (1986) Species conservation and systematics: The dilemma of the subspecies. Trends Ecol Evol 1:9–10

Roberge JM, Angelstam P (2004) Usefulness of the umbrella species concept as a conservation tool. Conserv Biol 18:76–85. https://doi.org/10.1111/j.1523-1739.2004.00450.x

Robledo D, Palaiokostas C, Bargelloni L, *et al* (2018) Applications of genotyping by sequencing in aquaculture breeding and genetics. Rev Aquac 10:670–682. https://doi.org/10.1111/raq.12193

Rochette NC, Catchen JM (2017) Deriving genotypes from RAD-seq short-read data using Stacks. Nat Protoc 12:2640–2659. https://doi.org/10.1038/nprot.2017.123

Rochette NC, Rivera-Colón AG, Catchen JM (2019) Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. Mol Ecol 28:4737–4754. https://doi.org/10.1111/mec.15253

Rousset F (2008) GENEPOP'007: A complete re-implementation of the GENEPOP software for Windows and Linux. Mol Ecol Resour 8:103–106. https://doi.org/10.1111/j.1471-8286.2007.01931.x

Rozas-Serri M, Peña A, Maldonado L (2018) Transcriptomic profiles of post-smolt Atlantic salmon challenged with

Piscirickettsia salmonis reveal a strategy to evade the adaptive immune response and modify cell-autonomous immunity. Dev Comp Immunol 81:348–362. https://doi.org/10.1016/j.dci.2017.12.023

Ruzzante DE, Hansen MM, Meldrup D (2001) Distribution of individual inbreeding coefficients, relatedness and influence of stocking on native anadromous brown trout (*Salmo trutta*) population structure. Mol Ecol 10:2107–2128. https://doi.org/10.1046/j.1365-294X.2001.01352.x

Ryman N, Laikre L, Hössjer O (2019) Do estimates of contemporary effective population size tell us what we want to know? Mol Ecol 28:1904–1918. https://doi.org/10.1111/mec.15027

Saint-Pé K, Blanchet S, Tissot L, *et al* (2018) Genetic admixture between captive-bred and wild individuals affects patterns of dispersal in a brown trout (*Salmo trutta*) population. Conserv Genet 19:1269–1279. https://doi.org/10.1007/s10592-018-1095-2

Saint-Pé K, Leitwein M, Tissot L, *et al* (2019) Development of a large SNPs resource and a low-density SNP array for brown trout (*Salmo trutta*) population genetics. BMC Genomics 20:582. https://doi.org/10.1186/s12864-019-5958-9

Saltveit SJ (2006) The effects of stocking Atlantic salmon, *Salmo salar*, in a Norwegian regulated river. Fish Manag Ecol 13:197–205. https://doi.org/10.1111/j.1365-2400.2006.00494.x

Sánchez-Hernández J, Servia MJ, Vieira-Lanero R, Cobo F (2013) Ontogenetic dietary shifts in a predatory freshwater fish species: the brown trout as an example of a dynamic fish species. In: Türker H (ed) New advances and contributions to fish biology. InTech, Croatia, pp 271–298

Sánchez-Hernández J (2016) Do age-related changes in feeding habits of brown trout alter structural properties of food webs? Aquat Ecol 50:685–695 https://doi.org/10.1007/s10452-016-9586-z

Sankaraman S, Sridhar S, Kimmel G, *et al* (2008) Estimating local ancestry in admixed populations. Am J Hum Genet 82(2):290–303. https://doi.org/10.1016/j.ajhg.2007.09.022

Sanz N, García-Marín JL, Pla C (2002) Managing fish populations under mosaic relationships. The case of brown trout (*Salmo trutta*) in peripheral Mediterranean populations. Conserv Genet 3:385–400. https://doi.org/10.1023/A:1020527420654

Sanz N, Cortey M, Pla C, García-Marín JL (2006) Hatchery introgression blurs ancient hybridization between brown trout (*Salmo trutta*) lineages as indicated by complementary allozymes and mtDNA markers. Biol Conserv 130:278–289. https://doi.org/10.1016/j.biocon.2005.12.023

Sanz N, Araguas RM, Fernández R, *et al* (2009) Efficiency of markers and methods for detecting hybrids and introgression in stocked populations. Conserv Genet 10:225–236. https://doi.org/10.1007/s10592-008-9550-0

Sanz, N (2018) Phylogeographic History of brown trout: A Review. In: Lobón-Cerviá J, Sanz N (eds) Brown trout: Biology, ecology and management. Wiley, West Sussex pp 17–64

Sanz N, Araguas RM, Fernández-Cebrián R, Lobón-Cerviá J (2019) Factors modelling population structure in brown trout *Salmo trutta* L.: genetic monitoring of populations in Esva River (northwestern Spain). Hydrobiologia 837:117–131. https://doi.org/10.1007/s10750-019-3965-0

Sanz N, Nebot A, Araguas RM (2020) Microsatellites as a good approach for detecting triploidy in brown trout hatchery stocks. Aquaculture 523: 735218. https://doi.org/10.1016/j.aquaculture.2020.735218

Schmidt H, Bernet D, Wahli T, *et al* (1999) Active biomonitoring with brown trout and rainbow trout in diluted sewage plant effluents. J Fish Biol 54:585–596 https://doi.org/10.1006/jfbi.1998.0892

Schöffmann J, Sušnik S, Snoj A (2007) Phylogenetic origin of *Salmo trutta* L 1758 from Sicily, based on mitochondrial and nuclear DNA analyses. Hydrobiologia 575:51–55. https://doi.org/10.1007/s10750-006-0281-2

Schultz K (2003) Ken Schultz's Field Guide to Freshwater Fish. Wiley, New York

Schuster SC (2008) Next-generation sequencing transforms today's biology. Nat. Methods 5:16–18. https://doi.org/10.1038/nmeth1156

Shafer ABA, Wolf JBW, Alves PC, *et al* (2015) Genomics and the challenging translation into conservation practice. Trends Ecol Evol 30(2):78–87. https://doi.org/10.1016/j.tree.2014.11.009

Shafer ABA, Peart CR, Tusso S, *et al* (2017) Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. Methods Ecol Evol 8:907–917. https://doi.org/10.1111/2041-210X.12700

Škraba Jurlina D, Marić A, Karanović J, *et al* (2018) Effect of the introgression of Atlantic brown trout, *Salmo trutta*, into Adriatic trout, *Salmo farioides* in a stream at the drainage area of the Adriatic Sea basin of Montenegro. Acta Ichthyol Piscat 48(4):363–372. https://doi.org/10.3750/AIEP/02491

Slatkin M, Voelm L (1991) F_{ST} in a hierarchical island model. Genetics 127(3):627–629. https://doi.org/10.1093/genetics/127.3.627

Slatko BE, Gardner AF, Ausubel FM (2018) Overview of Next-Generation Sequencing Technologies. Curr Protoc Mol Biol 122:e59. https://doi.org/10.1002/cpmb.59

Solà C, Ordeix M, Pou-Rovira Q, *et al* (2011) Longitudinal connectivity in hydromorphological quality assessments of rivers. The ICF index: A river connectivity index and its application to Catalan rivers. Limnetica 30(2): 273–292 https://doi.org/10.23818/limn.30.21

Sorensen PW, Cardwell J, Essington T, *et al* (1995) Reproductive interactions between sympatric brook and brown

trout in a small Minnesota stream. Can J Fish Aquat Sci 52:1958–1965. https://doi.org/10.1139/f95-787

Sovic MG, Fries AC, Gibbs HL (2015) AftrRAD: a pipeline for accurate and efficient *de novo* assembly of RADseq data. Mol Ecol Resour 15:1163–1171. https://doi.org/10.1111/1755-0998.12378

Splendiani A, Palmas F, Sabatini A, *et al* (2019) The name of the trout: considerations on the taxonomic status of the *Salmo trutta* L., 1758 complex (Osteichthyes: Salmonidae) in Italy. Eur Zool J 86(1):432–442. https://doi.org/10.1080/24750263.2019.1686544

Storz JF (2016) Gene duplication and evolutionary innovations in hemoglobin-oxygen transport. Physiology 31(3):223–232. https://doi.org/10.1152/physiol.00060.2015

Stanley RRE, Jeffery NW, Wringe BF, *et al* (2017) *genepopedit*: a simple and flexible tool for manipulating multilocus molecular data in R. Mol Ecol Resour 17:12–18. https://doi.org/10.1111/1755-0998.12569

Sudhagar A, El-Matbouli M, Kumar G (2020) Identification and expression profiling of toll-like receptors of brown trout (*Salmo trutta*) during proliferative kidney disease. Int J Mol Sci 21(11):3755. https://doi.org/10.3390/ijms21113755

Sušnik S, Schöffmann J, Weiss S (2005) Genetic verification of native brown trout from the Persian Gulf (Catak Cay River, Tigris basin). J Fish Biol 67:879–884. https://doi.org/10.1111/j.0022-1112.2005.00780.x

Sušnik S, Pustovrh G, Jesenšek D, Snoj A (2015) Population genetic SNP analysis of marble and brown trout in a hybridization zone of the Adriatic watershed in Slovenia. Biol Conserv 184:239–250. https://doi.org/10.1016/j.biocon.2015.01.033

Takeuchi T, Koyanagi R, Gyoja F, *et al* (2016) Bivalve-specific gene expansion in the pearl oyster genome: implications of adaptation to a sessile lifestyle. Zool Lett 2: 3. https://doi.org/10.1186/s40851-016-0039-2

Torkamaneh D, Laroche J, Belzile F (2016) Genome-wide SNP calling from genotyping by sequencing (GBS) data: A comparison of seven pipelines and two sequencing technologies. PLoS ONE 11(8): e0161333. https://doi.org/10.1371/journal.pone.0161333

Torkamaneh D, Laroche J, Bastien M, *et al* (2017) Fast-GBS: A new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. BMC Bioinformatics 18:5. https://doi.org/10.1186/s12859-016-1431-9

Tougard C, Justy F, Guinand B, *et al* (2018) *Salmo macrostigma* (Teleostei, Salmonidae): Nothing more than a brown trout (*S. trutta*) lineage? J Fish Biol 93:302–310. https://doi.org/10.1111/jfb.13751

Turan D, Doğan E, Kaya C, Kanyılmaz M (2014) *Salmo kottelati*, a new species of trout from Alakır Stream, draining to the Mediterranean in southern Anatolia, Turkey (Teleostei, Salmonidae). Zookeys 462:135–151. https://doi.org/10.3897/zookeys.462.8177

Valette T, Leitwein M, Lascaux J, *et al* (2020). Spotting genome-wide pigmentation variation in a brown trout admixture context. bioRxiv. https://doi.org/10.1101/2020.07.23.217109

Vera M, Cortey M, Sanz N, García-Marín JL (2010) Maintenance of an endemic lineage of brown trout (*Salmo trutta*) within the Duero river basin. J Zool Syst Evol Res 48:181–187. https://doi.org/10.1111/j.1439-0469.2009.00547.x

Vera M, Sourinejad I, Bouza C, *et al* (2011) Phylogeography, genetic structure, and conservation of the endangered Caspian brown trout, *Salmo trutta caspius* (Kessler, 1877), from Iran. Hydrobiologia 664:51–67. https://doi.org/10.1007/s10750-010-0581-4

Vera M, García-Marín JL, Martínez P, *et al* (2013) Identification and conservation of remnant genetic resources of brown trout in relict populations from Western Mediterranean streams. Hydrobiologia 707:29–45. https://doi.org/10.1007/s10750-012-1402-8

Vera M, García-Marín JL, Martínez P, Bouza C (2015) Phylogenetic diversity within the endemic brown trout Duero lineage: Implications for conservation and management. Mar Freshw Res 66(11):1066–1071. https://doi.org/10.1071/MF14286

Vera M, Carlsson J, Carlsson J El, *et al* (2016) Current genetic status, temporal stability and structure of the remnant wild European flat oyster populations: conservation and restoring implications. Mar Biol 163:239. https://doi.org/10.1007/s00227-016-3012-x

Vera M, Martinez P, Bouza C (2018) Stocking impact, population structure and conservation of wild brown trout populations in inner Galicia (NW Spain), an unstable hydrologic region. Aquat Conserv Mar Freshw Ecosyst 28:435–443. https://doi.org/10.1002/aqc.2856

Vilas R, Bouza C, Castro J, *et al* (2010) Management units of brown trout from Galicia (NW: Spain) based on spatial genetic structure analysis. Conserv Genet 11:897–906. https://doi.org/10.1007/s10592-009-9934-9

Vøllestad LA, Lillehammer T (2000) Individual variation in early life-history traits in brown trout. Ecol Freshw Fish 9:242–247. https://doi.org/10.1111/j.1600-0633.2000.eff090407.x

von Bubnoff A (2008) Next-Generation Sequencing: The Race Is On. Cell 132(5):721–723. https://doi.org/10.1016/j.cell.2008.02.028

Wang J (2016) A comparison of single-sample estimators of effective population sizes from genetic marker data. Mol Ecol 25:4692–4711. https://doi.org/10.1111/mec.13725

Wang S, Hard JJ, Utter F (2002) Salmonid inbreeding: A review. Rev Fish Biol Fish 11:301–319. https://doi.org/10.1023/A:1021330500365

Wang S, Meyer E, Mckay JK, *et al* (2012) 2b-RAD: A simple and flexible method for genome-wide genotyping. Nat Methods 9:808–810. https://doi.org/10.1038/nmeth.2023

Wang X, Chen L, Ma J (2019) Genomic introgression through interspecific hybridization counteracts genetic bottleneck

during soybean domestication. Genome Biol 20:1–15. https://doi.org/10.1186/s13059-019-1631-5

Wang Y, Guo X (2004) Chromosomal rearrangement in pectinidae revealed by rRNA loci and implications for bivalve evolution. Biol Bull 207(3):247–256. https://doi.org/10.2307/1543213

Waples RS (2006) A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. Conserv Genet 7:167. https://doi.org/10.1007/s10592-005-9100-y

Waples RS, Naish KA, Primmer CR (2020) Conservation and management of salmon in the age of genomics. Annu Rev Anim Biosci 8:117–143. https://doi.org/10.1146/annurev-animal-021419-083617

Webster TM, Bury N, Van Aerle R, *et al* (2013) Global transcriptome profiling reveals molecular mechanisms of metal tolerance in a chronically exposed wild population of brown trout. Environ Sci Technol 47(15): 8869–8877. https://doi.org/10.1021/es401380p

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x

Weiss S, Ferrand N (2010) Current perspectives in phylogeography and the significance of South European refugia in the creation and maintenance of European biodiversity. In: Weiss

S, Ferrand N (eds) Phylogeography Of Southern European Refugia. Springer, Dordrecht, pp 341–357

Welcomme R (1988) International introductions of inland aquatic species. FAO Fisheries Technical Paper No. 294. Food and Agriculture Organization of the United Nations (FAO), Rome

Wetterstrand KA. DNA Sequencing Costs: Data | NHGRI. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data. Accessed 1 Jul 2020

Wohl E (2017) Connectivity in rivers. Prog Phys Geogr. 41(3):345-362. https://doi.org/10.1177/0309133317714972

Wong K, Zhang J, Yan S (2019) DNA sequencing technologies: Sequencing data protocols and bioinformatics tools. 52(5):98. https://doi.org/10.1145/3340286

Wright B, Farquharson KA, McLennan EA, *et al* (2019) From reference genomes to population genomics: Comparing three reference-aligned reduced-representation sequencing pipelines in two wildlife species. BMC Genomics 20:453. https://doi.org/10.1186/s12864-019-5806-y

Wright S (1951) The genetic structure of populations Annals of Eugenics 15:323–354

Wright S (1969) Evolution and the Genetics of Populations, vol.2, The Theory of Gene Frequencies. Univ. Chicago Press, Chicago

Yan X, Nie H, Huo Z, *et al* (2019) Clam genome sequence clarifies the molecular basis of its benthic adaptation and

extraordinary shell color diversity. iScience. https://doi.org/10.1016/j.isci.2019.08.049

Yano A, Guyomard R, Nicol B, *et al* (2012) An immune-related gene evolved into the master sex-determining gene in rainbow trout, *Oncorhynchus mykiss*. Curr Biol 22(15):1423–1428. https://doi.org/10.1016/j.cub.2012.05.045

Yano A, Nicol B, Jouanno E, *et al* (2013) The sexually dimorphic on the Y-chromosome gene (*sdY*) is a conserved male-specific Y-chromosome sequence in many salmonids. Evol Appl 6:486–496. https://doi.org/10.1111/eva.12032

Zhang J, Chiodini R, Badr A, Zhang G (2011) The impact of next-generation sequencing on genomics. J Genet Genomics 38(3):95–109. https://doi.org/10.1016/j.jgg.2011.02.003

Zueva KJ, Lumme J, Veselov AE, *et al* (2018) Genomic signatures of parasite-driven natural selection in north European Atlantic salmon (*Salmo salar*). Mar Genomics 39:26–38. https://doi.org/10.1016/j.margen.2018.01.001

SUPPLEMENTARY TABLES

Table S1. Spatial coordinates of sampling localities for the five species used in Chapter 3. Coordinates are in WGS 84 Datum format.

Species	Country	Location	Coordinates	Number of samples
Manila clam				
	Italy	Chioggia	45.239, 12.298	30
	Italy	Porto Marghera	45.462, 12.277	30
	Italy	Po River mouth	44.954, 12.450	25
	Spain	Vigo	42.156, -8.844	25
Common cockle				
	France	Somme Bay	50.244, 1.574	30
	Portugal	Ría Formosa	36.997, -7.830	30
	Spain	Campelo	42.420, -8.684	30
	Spain	Miño	43.361, -8.205	30
Brown trout				
	Spain	Águeda River	40.325, -6.763	15
	Spain	Omaña River	42.787, -6.043	20
	Spain	Pisuerga River	42.782, -4.258	17
Silver catfish				
	Uruguay	Sauce Lagoon	-34.830, -55.059	10
	Uruguay	Uruguay River Basin	-32.188, -57.628 -31.197, -57.165	11
Small-spotted catshark				
		Irish Sea	53.916, -5.229	15
		North Sea	57.125, -0.602	13

Table S2A. Stacks 2 process_radtags options used to filtering reads by quality criteria.

Module	Species	Main options
process_radtags	Small-spotted catshark	-w 0.1 -s 20
process_radtags	Brown trout	-w 0.25 -s 30
process_radtags	Manila clam, common cockle, silver catfish	-w 0.25 -s 20

Table S2B. Stacks 2 main options used in this Thesis. With Stacks 2 *de novo* and reference genome approaches were used.

Module	Species	Approach	Main options
ustacks	Mollusc	*de novo*	-m 3 -M 3 -N 0 --disable-gapped -d --model_type snp --alpha 0.05
ustacks	Fishes	*de novo*	-m 3 -M 2 -N 0 --disable-gapped -d --model_type snp --alpha 0.05
cstacks	Mollusc	*de novo*	--disable_gapped -n 3
cstacks	Fishes	*de novo*	--disable_gapped -n 2
sstacks	All	*de novo*	--disable_gapped
tsv2bam	All	*de novo*	-
gstacks	All	*de novo* and Reference genome	--model marukilow var_alpha: 0.05 gt_alpha: 0.05
populations	All	*de novo* and Reference genome	-

Table S2C. Meyer's 2b-RAD v2.1 pipeline main options used in Chapter 3.

To prepare the reference Option 2 was used, a *de novo* reference by clustering reads (see guide for v3.0; http://eli-meyer.github.io/2bRAD_utilities/). The combined dataset used to prepare reference was about 20 million filtered reads in all species.

Steps from guide	Script/program	Species	Main options
Prepare reference			
	BuildRef.pl	All	qthd = 30
	cd-hit-est	All	-b 1
	cd-hit-est	Mollusc	-c = 0.916
	cd-hit-est	Small-spotted catshark	-c = 0.937
	cd-hit-est	Brown trout and Silver catfish	-c = 0.944
Align reads against reference			
	Bowtie 1.1.2	Mollusc	-v 3
		Fishes	-v 2
Determine genotypes from alignments			
	NFGenotyper.pl	All	mincov = 3
		All	Nucleotide frequencies: 0.1-0.2

Table S3. Number of SNPs and population metrics for Manila clam (*Ruditapes philippinarum*) samples (N = 110, four localities). In dark grey, the results obtained using either a reference genome approach (RG) or shared SNPs between RG and STA (RG-STA) or ALT (RG-ALT) de novo SNP panels. Mean ± SD across loci and populations of observed heterozygosity (H_o), expected heterozygosity (H_e), allelic richness (A_r), inbreeding coefficient (F_{IS}); Hardy-Weinberg *p*-value per population and percentage of loci in Hardy-Weinberg disequilibrium per population (*p*-value < 0.05); global interpopulation fixation index (Global F_{ST}); population structure units detected (STR groups); number of outliers detected with BayeScan.

Filtering steps				Population metrics							
Panels	Initial SNP number	5th step SNP number	Final SNP number	H_o (± SD), H_e (± SD)	A_r (±SD)	F_{IS} (±SD)	HW (global)	HW (%P < 0.05)	Global F_{ST}	STR (Groups)	Outliers log10(P)>2 [>1.5]
STA	209,059	825	479	0.120±0.014, 0.163±0.005	1.698±0.010	0.237± 0.054	0.000 0.000 0.000 0.000	31.8 25.1 20.2 26.7	0.006	No	1[1]
ALT	215,042	1,565	956	0.103±0.005, 0.135±0.006	1.660±0.024	0.251± 0.034	0.000 0.000 0.000 0.000	32.7 30.4 25.0 28.7	0.003	No	0[0]
COM	-	-	206	0.138±0.017, 0.170±0.008	1.713±0.019	0.195± 0.058	0.000 0.000 0.000 0.000	26.0 20.7 18.4 28.8	0.004	No	0[0]
MER	-	-	1,084	0.108±0.010, 0.140±0.000	1.668±0.015	0.228± 0.044	0.000 0.000 0.000 0.000	31.6 29.4 23.2 26.6	0.005	No	1[1]
RG	122,334	512	289	0.123±0.013, 0.158±0.010	1.703±0.046	0.222± 0.050	0.000 0.000 0.000 0.000	26.4 29.2 23.0 24.0	0.006	No	0[0]
RG-STA	-	-	138	0.133±0.013, 0.173±0.005	1.720±0.024	0.206± 0.054	0.000 0.000 0.000 0.000	29.9 27.6 24.2 24.7	0.003	No	0[0]
RG-ALT	-	-	109	0.140±0.014, 0.173±0.010	1.723±0.036	0.202± 0.046	0.000 0.000 0.000 0.000	24.6 31.7 19.3 26.8	0.006	No	0[0]

Table S4. Number of SNPs and population metrics for common edible cockle (*Cerastoderma edule*) samples (N = 120, four localities). Mean ± SD across loci and populations of observed heterozygosity (H_o), expected heterozygosity (H_e), allelic richness (A_r), inbreeding coefficient (F_{IS}); Hardy-Weinberg *p*-value per population and percentage of loci in Hardy-Weinberg disequilibrium per population (*p*-value < 0.05); global interpopulation fixation index (Global F_{ST}); population structure units detected (STR groups); number of outliers detected with BayeScan.

Filtering steps				Population metrics							
Panels	Initial SNP number	5th step SNP number	Final SNP number	H_o (± SD), H_e (± SD)	A_r (±SD)	F_{IS} (±SD)	HW (global)	HW (%P < 0.05)	Global F_{ST}	STR (Groups)	Outliers log10(P)>2 [>1.5]
STA	356,389	3,299	2,218	0.145±0.010, 0.157±0.005	1.707±0.026	0.086± 0.028	0.000 0.900 0.000 0.000	14.4 10.7 16.6 15.3	0.033	Yes (3)	14[18]
ALT	426,317	4,253	2,990	0.125±0.006, 0.140±0.000	1.683±0.024	0.120± 0.026	0.000 0.000 0.000 0.000	18.0 13.8 15.9 17.9	0.029	Yes (3)	16[22]
COM	-	-	1,185	0.150±0.008, 0.160±0.000	1.733±0.024	0.066± 0.036	0.400 1.000 0.088 0.000	11.0 6.3 11.6 12.2	0.032	Yes (3)	7[10]
MER	-	-	3,725	0.133±0.005, 0.150±0.000	1.690±0.022	0.114± 0.029	0.000 0.000 0.000 0.000	17.2 13.3 17.6 17.9	0.030	Yes (3)	26[18]

Table S5. Number of SNPs and population metrics for brown trout (*Salmo trutta*) samples (N = 52, three localities). In dark grey, the results obtained using either a reference genome approach (RG) or shared SNPs between RG and STA (RG-STA) or ALT (RG-ALT) de novo SNP panels. Mean ± SD across loci and populations of observed heterozygosity (H_o), expected heterozygosity (H_e), allelic richness (A_r), inbreeding coefficient (F_{IS}); Hardy-Weinberg *p*-value per population and percentage of loci in Hardy-Weinberg disequilibrium per population (*p*-value < 0.05); global interpopulation fixation index (Global F_{ST}); population structure units detected (STR groups); number of outliers detected with BayeScan.

Filtering steps				Population metrics							
Panels	Initial SNP number	5th step SNP number	Final SNP number	H_o (± SD), H_e (± SD)	A_r (±SD)	F_{IS} (±SD)	HW (global)	HW (%P < 0.05)	Global F_{ST}	STR (Groups)	Outliers log10(P)>2 [>1.5]
STA	56,074	18,793	6,446	0.243±0.023, 0.190±0.017	1.523±0.041	-0.269± 0.023	0.000 0.000 0.000	13.9 16.5 16.8	0.376	Yes (2-3)	0[0]
ALT	287,503	21,511	10,080	0.250±0.035, 0.187±0.021	1.520±0.046	-0.336± 0.028	0.000 0.000 0.000	15.8 14.3 14.0	0.370	Yes (2-3)	0[0]
COM	-	-	4,792	0.200±0.026, 0.170±0.026	1.470±0.044	-0.179± 0.038	0.000 0.987 0.000	12.3 9.1 11.8	0.442	Yes (2-3)	0[0]
MER	-	-	11,305	0.257±0.029, 0.193±0.023	1.533±0.042	-0.333± 0.024	0.000 0.000 0.000	15.8 16.6 16.4	0.348	Yes (2-3)	0[4]
RG	41,972	5,178	4,104	0.180±0.010, 0.163±0.023	1.453±0.042	-0.125± 0.048	1.000 1.000 1.000	4.3 3.6 7.2	0.500	Yes (2-3)	0[0]
RG-STA	-	-	3,000	0.180±0.020, 0.160±0.017	1.450±0.044	-0.115± 0.052	1.000 1.000 1.000	4.0 3.4 6.8	0.505	Yes (2-3)	0[0]
RG-ALT	-	-	3,324	0.177±0.020, 0.160±0.017	1.453±0.042	-0.113± 0.054	1.000 1.000 1.000	3.7 3.2 6.8	0.504	Yes (2-3)	0[0]

Table S6. Number of SNPs and population metrics for silver catfish (*Rhamdia quelen*) samples (N = 21, two localities). Mean ± SD across loci and populations of observed heterozygosity (H_o), expected heterozygosity (H_e), allelic richness (A_r), inbreeding coefficient (F_{IS}); Hardy-Weinberg *p*-value per population and percentage of loci in Hardy-Weinberg disequilibrium per population (*p*-value < 0.05); global interpopulation fixation index (Global F_{ST}); population structure units detected (STR groups); number of outliers detected with BayeScan.

Filtering steps				Population metrics							
Panels	Initial SNP number	5th step SNP number	Final SNP number	H_o (± SD), H_e (± SD)	A_r (±SD)	F_{IS} (±SD)	HW (global)	HW (%P < 0.05)	Global F_{ST}	STR (Groups)	Outliers log10(P)>2 [>1.5]
STA	104,656	27,769	21,468	0.235±0.049, 0.230±0.056	1.690±0.180	-0.004± 0.032	1.000 1.000	4.0 4.6	0.452	Yes (2)	0[0]
ALT	125,823	27,508	22,481	0.235±0.049, 0.230±0.057	1.680±0.170	-0.012± 0.036	1.000 1.000	5.1 4.9	0.453	Yes (2)	0[0]
COM	-	-	17,459	0.230±0.057, 0.230±0.057	1.685±0.177	0.002± 0.024	1.000 1.000	3.2 3.9	0.465	Yes (2)	0[0]
MER	-	-	25,042	0.235±0.049, 0.235±0.049	1.690±0.170	-0.014± 0.038	1.000 1.000	4.9 5.1	0.451	Yes (2)	0[0]

Table S7. Number of SNPs and population metrics for small-spotted catshark (*Scyliorhinus canicula*) samples (N = 28, two localities). Mean ± SD across loci and populations of observed heterozygosity (H_o), expected heterozygosity (H_e), allelic richness (A_r), inbreeding coefficient (F_{IS}); Hardy-Weinberg *p*-value per population and percentage of loci in Hardy-Weinberg disequilibrium per population (*p*-value < 0.05); global interpopulation fixation index (Global F_{ST}); population structure units detected (STR groups); number of outliers detected with BayeScan.

Filtering steps				Population metrics							
Panels	Initial SNP number	5th step SNP number	Final SNP number	H_o (± SD), H_e (± SD)	A_r (±SD)	F_{IS} (±SD)	HW (global)	HW (%P < 0.05)	Global F_{ST}	STR (Groups)	Outliers log10(P)>2 [>1.5]
STA	163,088	1,817	913	0.545±0.021, 0.355±0.007	1.925±0.021	-0.528± 0.035	0.000 0.000	27.6 8.4	0.002	No	0[0]
ALT	417,292	5,446	3,286	0.520±0.000, 0.340±0.000	1.905±0.007	-0.541± 0.018	0.000 0.000	22.8 9.1	0.002	No	0[0]
COM	-	-	218	0.460±0.028, 0.325±0.021	1.915±0.021	-0.406± 0.024	0.000 0.984	15.7 5.4	0.004	No	0[0]
MER	-	-	3,697	0.535±0.007, 0.340±0.000	1.915±0.007	-0.544± 0.020	0.000 0.000	23.9 9.3	0.002	No	0[0]

Table S8. SNPs with F_{ST} = 1 between wild and hatchery individuals. These SNPs belong to the smallest panel (N = 9) used to identify hatchery ancestry. SNP ID codes are those used in the Table 7.

SNP ID (Table 7)	RefSeq sequence (Chr.)	Genome position
1	NC_042959.1 (3)	12,826,411
2	NC_042959.1 (3)	14,460,115
3	NC_042959.1 (3)	15,524,355
4	NC_042960.1 (4)	20,745,779
5	NC_042964.1 (8)	15,577,318
6	NC_042965.1 (9)	27,603,192
7	NC_042970.1 (14)	19,329,385
8	NC_042970.1 (14)	22,593,032
9	NC_042970.1 (14)	43,390,010

Table S9. Consistent outliers detected through the four analyses with different subset of samples performed. The number of analyses in which each SNP was identified as a consistent outlier is indicated. ***AT and MED*** is the analysis with twelve populations; ***AT-S*** is the analysis with ten populations from Atlantic Slope; ***DU-B*** is the analysis with six populations from Duero Basin and ***MI-B*** the analysis with four populations from Miño Basin.

Chromosome	Position	Number of times detected	*AT and MED*	*AT-S*	*DU-B*	*MI-B*
1	5,448,642	1	–	–	1	–
1	51,421,298	3	1	1	1	–
1	51,939,424	1	–	–	1	–
1	56,669,167	1	–	–	1	–
1	60,403,198	1	–	1	–	–
2	27,052,157	1	–	1	–	–
2	44,089,029	1	–	–	1	–
2	53,975,176	1	–	1	–	–
2	54,776,786	1	–	1	–	–
2	55,611,146	1	–	1	–	–
3	28,594,860	1	–	1	–	–
3	32,608,808	1	–	1	–	–
3	54,956,067	1	–	1	–	–
3	56,007,797	1	–	–	1	–

Table S9. ***(Cont.)***

Chromosome	Position	Number of times detected	*AT and MED*	*AT-S*	*DU-B*	*MI-B*
3	67,749,091	1	–	1	–	–
4	2,283,920	1	–	–	1	–
4	24,387,913	1	–	–	1	–
4	30,358,595	1	–	1	–	–
4	31,196,560	2	1	1	–	–
4	32,671,226	1	–	–	1	–
4	32,810,939	2	–	1	1	–
4	33,321,232	1	–	–	1	–
4	36,175,811	1	–	1	–	–
4	47,623,400	1	–	–	1	–
4	51,579,033	1	–	1	–	–
4	55,256,715	1	–	1	–	–
4	57,718,518	2	1	1	–	–
4	70,118,763	1	–	–	1	–
6	2,569,032	2	1	1	–	–
6	17,065,846	1	–	1	–	–

Table S9. ***(Cont.)***

Chromosome	Position	Number of times detected	*AT and MED*	*AT-S*	*DU-B*	*MI-B*
6	26,357,716	1	1	–	–	–
7	2,738,846	2	1	1	–	–
7	3,852,789	1	–	–	1	–
7	16,224,119	1	–	–	1	–
7	21,897,148	2	1	1	–	–
7	34,603,472	1	–	–	1	–
7	51,796,043	1	–	1	–	–
7	52,415,927	3	1	1	1	–
9	29,989,557	1	–	–	1	–
9	43,848,589	1	–	–	1	–
10	15,471,120	1	–	1	–	–
10	24,487,806	3	1	1	1	–
12	37,238,555	1	–	–	1	–
12	68,780,493	2	1	1	–	–
13	3,084,536	2	–	1	1	–
13	17,054,719	1	–	1	–	–

Table S9. ***(Cont.)***

Chromosome	Position	Number of times detected	*AT and MED*	*AT-S*	*DU-B*	*MI-B*
13	50,422,333	1	–	1	–	–
13	56,916,679	1	–	–	1	–
13	57,825,984	2	–	1	1	–
14	332,965	1	–	–	1	–
14	476,020	1	–	1	–	–
14	24,608,651	1	–	1	–	–
14	26,492,247	2	–	1	1	–
14	28,276,551	1	–	–	1	–
14	29,533,487	3	1	1	1	–
14	31,786,048	1	–	1	–	–
14	36,777,948	2	–	1	1	–
14	42,150,209	1	–	–	1	–
14	66,355,767	1	–	–	1	–
14	70,275,783	2	1	1	–	–
14	78,924,480	2	1	1	–	–

Table S9. ***(Cont.)***

Chromosome	Position	Number of times detected	*AT and MED*	*AT-S*	*DU-B*	*MI-B*
15	14,812,814	1	–	1	–	–
15	20,895,905	1	–	1	–	–
15	31,790,405	1	1	–	–	–
15	39,749,930	2	–	1	1	–
15	42,271,646	1	–	–	1	–
15	47,866,615	1	–	1	–	–
16	3,209,269	2	–	1	1	–
17	23,350,408	1	–	1	–	–
17	30,233,362	1	–	1	–	–
17	39,541,192	1	–	–	1	–
17	51,689,193	2	1	1	–	–
18	4,952,664	1	–	1	–	–
18	15,178,965	1	–	–	1	–
18	50,307,805	1	–	1	–	–
19	9,857,462	1	–	–	1	–

Table S9. ***(Cont.)***

Chromosome	Position	Number of times detected	*AT and MED*	*AT-S*	*DU-B*	*MI-B*
19	27,096,514	3	1	1	1	–
20	15,618,457	1	–	–	1	–
20	18,837,175	1	–	1	–	–
20	40,004,007	2	1	1	–	–
20	45,967,814	2	1	1	–	–
21	1,239,810	2	1	1	–	–
21	1,511,014	1	–	1	–	–
21	31,113,584	1	–	1	–	–
21	50,554,172	3	1	1	1	–
22	1,035,523	1	–	1	–	–
22	2,308,846	1	–	–	1	–
22	2,518,313	1	–	–	1	–
22	22,181,155	1	–	–	–	1
22	38,240,039	1	–	–	1	–
22	41,582,124	1	–	–	1	–
23	4,703,828	1	–	–	1	–

Table S9. ***(Cont.)***

Chromosome	Position	Number of times detected	*AT and MED*	*AT-S*	*DU-B*	*MI-B*
23	9,965,482	1	1	–	–	–
23	22,318,597	1	–	1	–	–
23	23,982,720	1	–	1	–	–
24	3,222,404	2	1	1	–	–
24	13,995,161	1	–	1	–	–
25	25,679,541	1	–	1	–	–
25	25,733,646	1	–	–	1	–
25	33,562,840	1	–	1	–	–
26	7,999,450	1	–	–	1	–
27	14,358,290	1	–	–	1	–
27	24,958,253	1	–	–	1	–
27	29,638,716	1	1	–	–	–
28	22,468,354	1	–	1	–	–
28	42,469,238	1	–	–	1	–
29	24,824,634	1	–	1	–	–

Table S9. ***(Cont.)***

Chromosome	Position	Number of times detected	*AT and MED*	*AT-S*	*DU-B*	*MI-B*
30	6,263,480	1	–	–	1	–
30	16,858,679	2	–	1	1	–
30	38,733,991	1	–	–	–	1
31	6,561,783	2	–	1	1	–
31	11,300,472	1	–	1	–	–
31	25,649,994	2	–	1	1	–
31	26,173,744	1	–	1	–	–
31	28,710,572	1	–	1	–	–
32	9,944,805	1	–	–	1	–
33	551,383	1	–	–	–	1
33	18,701,990	1	–	–	1	–
33	35,307,836	1	–	1	–	–
34	8,935,108	1	–	–	1	–
34	10,367,936	2	–	1	1	–
34	19,931,917	1	–	1	–	–
34	27,603,005	3	1	1	1	–

Table S9. ***(Cont.)***

Chromosome	Position	Number of times detected	*AT and MED*	*AT-S*	*DU-B*	*MI-B*
34	41,678,099	1	–	1	–	–
35	2,072,553	1	–	–	–	1
35	11,655,446	1	–	–	1	–
35	28,642,614	1	–	–	1	–
36	9,218,404	1	–	–	1	–
36	40,696,831	3	1	1	1	–
37	3,338,485	1	–	1	–	–
38	2,796,834	1	–	1	–	–
38	3,980,248	1	–	–	1	–
38	14,310,064	1	–	1	–	–
38	14,871,809	1	–	–	1	–
39	1,642,441	1	–	1	–	–
39	3,914,496	1	–	1	–	–
39	20,076,786	1	–	–	1	–
39	21,482,835	1	–	1	–	–

Table S9. ***(Cont.)***

Chromosome	Position	Number of times detected	*AT and MED*	*AT-S*	*DU-B*	*MI-B*
39	24,932,595	1	–	–	1	–
40	11,311,726	2	1	1	–	–

Table S10. Consistent outliers detected through the four analyses with different subset of samples performed. Sel., type of selection suggested by BayeScan and Arlequin; B (Balancing) D (Divergent). Loc., outlier location, IGR (Intergenic Region), UTR, (Untranslated Region), I (Intron), E (Exon) or Pseudogene. S/nS, Synonymous or not Synonymous changes, in parentheses the amino acids.

Chromosome	Position	Sel.	Alleles	Loc.	S/nS	Name	Description
1	5,448,642	B	G/C	IGR	–	–	–
1	51,421,298	B	C/A	I/5'UTR	–	*hbba2/hbaa2*	hemoglobin subunit beta-1-like/hemoglobin subunit alpha-like
1	51,939,424	D	G/A	I	–	*ENSSTUG00000014083*	vascular cell adhesion protein 1-like
1	56,669,167	B	G/A	I/E	nS (Gly>Asp)	*CYGB/prpsap1*	cytoglobin-2-like/phosphoribosyl pyrophosphate synthase-associated protein 1
1	60,403,198	B	C/T	IGR	–	–	–
2	27,052,157	D	G/A	IGR	–	–	–
2	44,089,029	D	C/T	I	–	*ENSSTUG00000037614*	zinc finger MIZ domain-containing protein 1-like
2	53,975,176	D	G/T	IGR	–	–	–
2	54,776,786	D	G/T	IGR	–	–	–
2	55,611,146	D	G/A	IGR	–	–	–
3	28,594,860	D	C/T	E	nS (Arg>Cys)	*stx3a*	syntaxin-3-like
3	32,608,808	D	C/T	I	–	*srpx2*	sushi-repeat containing protein X-linked 2
3	54,956,067	B	T/A	IGR	–	–	–
3	56,007,797	D	T/G	I	–	*arhgef2*	rho guanine nucleotide exchange factor 2-like
3	67,749,091	D	G/A	I	–	*CTNNB1*	catenin beta-1
4	2,283,920	D	A/T	IGR	–	–	–
4	24,387,913	D	C/T	I	–	*cramp1*	protein cramped-like

Table S10. ***(Cont.)***

Chromosome	Position	Sel.	Alleles	Loc.	S/nS	Name	Description
4	30,358,595	B	C/T	IGR	–	–	–
4	31,196,560	B	C/G	IGR	–	–	–
4	32,671,226	D	C/T	I	–	*sgip1a*	SH3-containing GRB2-like protein 3-interacting protein 1
4	32,810,939	B	C/T	IGR	–	–	–
4	33,321,232	D	A/T	I	–	*DDAH1*	N(G),N(G)-dimethylarginine dimethylaminohydrolase 1-like
4	36,175,811	D	C/T	I	–	*trappc8*	trafficking protein particle complex subunit 8-like
4	47,623,400	D	C/T	E	S (Leu)	*rfc4*	replication factor C (activator 1) 4
4	51,579,033	D	A/T	IGR	–	–	–
4	55,256,715	D	C/T	IGR	–	–	–
4	57,718,518	B	T/C	IGR	–	–	–
4	70,118,763	B	A/T	IGR	–	–	–
6	2,569,032	B	G/T	I	–	*plod2*	procollagen-lysine,2-oxoglutarate 5-dioxygenase 2-like
6	17,065,846	B	A/T	I	–	*hs6st1b*	heparan-sulfate 6-O-sulfotransferase 1-B
6	26,357,716	B	A/G	IGR	–	–	–
7	2,738,846	B	G/A	IGR	–	–	–
7	3,852,789	B	C/A	IGR	–	–	–
7	16,224,119	D	G/T	E	nS (Ala> Ser)	*hyal4*	hyaluronidase-4-like
7	21,897,148	B	G/T	I	–	*itga11b*	integrin alpha-11-like
7	34,603,472	B	G/C	IGR	–	–	–
7	51,796,043	D	C/T	I	–	*CDK17*	cyclin-dependent kinase 17-like
7	52,415,927	B	C/T	I	–	*parietopsin*	parietopsin
9	29,989,557	D	G/A	I	–	*zmat4a*	zinc finger matrin-type protein 4-like
9	43,848,589	B	C/G	I	–	*RASGEF1B*	ras-GEF domain-containing family member 1B-A-like

Table S10. ***(Cont.)***

Chromosome	Position	Sel.	Alleles	Loc.	S/nS	Name	Description
10	15,471,120	D	G/A	I/I	–	*glud1b/ENSSTUG00000040492*	glutamate dehydrogenase, mitochondrial-like/alpha-synuclein-like
10	24,487,806	D	G/A	IGR	–	–	–
12	37,238,555	D	A/G	I	–	*wt1b*	Wilms tumor protein homolog
12	68,780,493	B	C/T	I	–	*MED13L*	mediator complex subunit 13L
13	3,084,536	D	T/C	IGR	–	–	–
13	17,054,719	D	C/T	I	–	*fgf18a*	fibroblast growth factor 18-like
13	50,422,333	B	C/T	I	–	*arhgap35a*	rho GTPase-activating protein 35-like
13	56,916,679	D	G/A	I	–	*eif3k*	eukaryotic translation initiation factor 3 subunit K
13	57,825,984	D	C/T	IGR	–	–	–
14	332,965	D	C/T	IGR	–	–	–
14	476,020	D	G/C	I	–	*slc6a1b*	sodium- and chloride-dependent GABA transporter 1-like
14	24,608,651	D	A/C	I	–	*namptb*	nicotinamide phosphoribosyltransferase-like
14	26,492,247	B	G/A	I	–	*magi3b*	membrane-associated guanylate kinase, WW and PDZ domain-containing protein 3-like
14	28,276,551	B	G/A	IGR	–	–	–
14	29,533,487	B	A/C	IGR	–	–	–
14	31,786,048	B	G/T	I	–	*ZNF385A*	zinc finger protein 385A-like
14	36,777,948	B	G/A	IGR	–	–	–
14	42,150,209	D	C/T	IGR	–	–	–
14	66,355,767	D	T/G	I	–	*wipf2a*	WAS/WASL-interacting protein family member 2-like

Table S10. ***(Cont.)***

Chromosome	Position	Sel.	Alleles	Loc.	S/nS	Name	Description
14	70,275,783	B	C/T	I	–	*sh3pxd2aa*	SH3 and PX domain-containing protein 2A-like
14	78,924,480	B	A/T	I	–	*LOC115147484*	uncharacterized LOC115147484
15	14,812,814	D	C/T	I	–	*GABRB2*	gamma-aminobutyric acid receptor subunit beta-2
15	20,895,905	D	C/A	IGR	–	–	–
15	31,790,405	D	G/T	I	–	*wwc1*	protein KIBRA-like
15	39,749,930	D	C/A	Pseudogene	–	*ENSSTUG00000015840*	–
15	42,271,646	D	A/C	I	–	*eda*	ectodysplasin A
15	47,866,615	D	G/C	IGR	–	–	–
16	3,209,269	B	G/A	I	–	*cenpp*	centromere protein P
17	23,350,408	D	T/G	I	–	*zfpm1*	zinc finger protein ZFPM1-like
17	30,233,362	D	G/A	I	–	*FRMD5*	FERM domain-containing protein 5
17	39,541,192	B	T/C	I	–	*itpr2*	inositol 1,4,5-trisphosphate receptor type 2-like
17	51,689,193	B	C/A	I	–	*bpgm*	bisphosphoglycerate mutase-like
18	4,952,664	B	C/G	E	S (Leu)	*ENSSTUG00000019910*	–
18	15,178,965	B	T/G	IGR	–	–	–
18	50,307,805	B	A/G	I	–	*rps6ka2*	ribosomal protein S6 kinase alpha-2-like
19	9,857,462	D	G/C	IGR	–	–	–
19	27,096,514	B	C/T	I	–	*kctd7*	BTB/POZ domain-containing protein KCTD7
20	15,618,457	D	C/T	IGR	–	–	–
20	18,837,175	D	C/T	I	–	*erbb4b*	receptor tyrosine-protein kinase erbB-4-like

Table S10. ***(Cont.)***

Chromosome	Position	Sel.	Alleles	Loc.	S/nS	Name	Description
20	40,004,007	B	C/T	IGR	–	–	–
20	45,967,814	B	C/T	I	–	*ENSSTUG00000022542*	inactive dipeptidyl peptidase 10-like
21	1,239,810	D	C/T	I	–	*lrrc6*	leucine rich repeat containing 6
21	1,511,014	D	G/A	IGR	–	–	–
21	31,113,584	D	G/T	I	–	*ADGRL2*	adhesion G protein-coupled receptor L2-like
21	50,554,172	B	T/C	IGR	–	–	–
22	1,035,523	D	T/C	IGR	–	–	–
22	2,308,846	D	T/C	IGR	–	–	–
22	2,518,313	D	C/A	IGR	–	–	–
22	22,181,155	D	A/C	I	–	*bend5*	ATP/GTP binding protein like 4
22	38,240,039	D	A/G	I	–	*DDAH1*	dimethylarginine dimethylaminohydrolase 1
22	41,582,124	D	A/T	IGR	–	–	–
23	4,703,828	B	T/C	IGR	–	–	–
23	9,965,482	B	C/T	I	–	*naa25*	N-alpha-acetyltransferase 25, NatB auxiliary subunit-like
23	22,318,597	D	C/T	IGR	–	–	–
23	23,982,720	B	G/A	IGR	–	–	–
24	3,222,404	B	G/A	E	S (Glu)	*ENSSTUG00000004134*	–
24	13,995,161	D	G/A	E	S (Ile)	*lrp2a*	low-density lipoprotein receptor-related protein 2-like
25	25,679,541	D	A/G	IGR	–	–	–
25	25,733,646	D	G/A	I	–	*mnat1*	MNAT1 component of CDK activating kinase
25	33,562,840	B	T/G	IGR	–	–	–
26	7,999,450	B	A/C	IGR	–	–	–

Table S10. ***(Cont.)***

Chromosome	Position	Sel.	Alleles	Loc.	S/nS	Name	Description
27	14,358,290	D	G/C	E	nS (Thr>Arg)	*ankdd1b*	ankyrin repeat and death domain-containing protein 1B-like
27	24,958,253	B	C/A	IGR	–	–	–
27	29,638,716	B	G/A	I	–	*trpv4*	transient receptor potential cation channel subfamily V member 4-like
28	22,468,354	B	C/A	I	–	*ANKS1A*	ankyrin repeat and sterile alpha motif domain containing 1A
28	42,469,238	D	G/A	I	–	*slc2a1a*	solute carrier family 2, facilitated glucose transporter member 1-like
29	24,824,634	D	G/A	I	–	*fes*	tyrosine-protein kinase Fes/Fps-like
30	6,263,480	D	G/A	I	–	*klf15*	Krueppel-like factor 15
30	16,858,679	D	C/A	IGR	–	–	–
30	38,733,991	D	G/A	IGR	–	–	–
31	6,561,783	D	C/T	IGR	–	–	–
31	11,300,472	D	C/T	IGR	–	–	–
31	25,649,994	B	G/C	I	–	*pard3ab*	partitioning defective 3 homolog
31	26,173,744	D	T/G	IGR	–	–	–
31	28,710,572	D	C/G	I	–	*slc44a5a*	choline transporter-like protein 5-A
32	9,944,805	D	C/T	3'-UTR	–	*si:dkey-225f5.4*	uncharacterized LOC115171022
33	551,383	D	A/T	I	–	*wdhd1*	WD repeat and HMG-box DNA binding protein 1
33	18,701,990	D	C/T	IGR	–	–	–
33	35,307,836	D	C/T	I	–	*dcdc2b*	doublecortin domain-containing protein 2B-like
34	8,935,108	B	G/A	E	nS (Arg>Cys)	*ENSSTUG00000050668*	–
34	10,367,936	B	C/T	I	–	*pvrl2l*	poliovirus receptor homolog
34	19,931,917	D	C/T	IGR	–	–	–

Table S10. ***(Cont.)***

Chromosome	Position	Sel.	Alleles	Loc.	S/nS	Name	Description
34	27,603,005	B	C/T	I	–	*adgrb2*	adhesion G protein-coupled receptor B2
34	41,678,099	B	G/T	IGR	–	–	–
35	2,072,553	D	C/T	IGR	–	–	–
35	11,655,446	D	A/G	I	–	*nrxn3b*	neurexin-3b-like
35	28,642,614	B	A/G	I	–	*rnf8*	ring finger protein 8
36	9,218,404	B	A/T	IGR	–	–	–
36	40,696,831	D	A/C	I	–	*SLC9A3*	sodium/hydrogen exchanger 3-like
37	3,338,485	B	T/G	I	–	*ENSSTUG00000044101*	FH1/FH2 domain-containing protein 3-like
38	2,796,834	B	C/T	I/I	–	*ENSSTUG00000027113/ENSSTUG00000027095*	zinc finger protein 883-like/-
38	3,980,248	B	G/T	IGR	–	–	–
38	14,310,064	D	G/A	I	–	*rptor*	regulatory-associated protein of mTOR-like
38	14,871,809	D	C/T	IGR	–	–	–
39	1,642,441	D	C/T	IGR	–	–	–
39	3,914,496	D	C/T	IGR	–	–	–
39	20,076,786	B	C/T	IGR	–	–	–
39	21,482,835	B	A/T	I	–	*mgat5*	alpha-1,6-mannosylglycoprotein 6-beta-N-acetylglucosaminyltransferase A-like
39	24,932,595	D	C/T	IGR	–	–	–
40	11,311,726	B	G/T	E	S (Arg)	*zgc:113263*	uncharacterized LOC115180156

SUPPLEMENTARY FIGURES

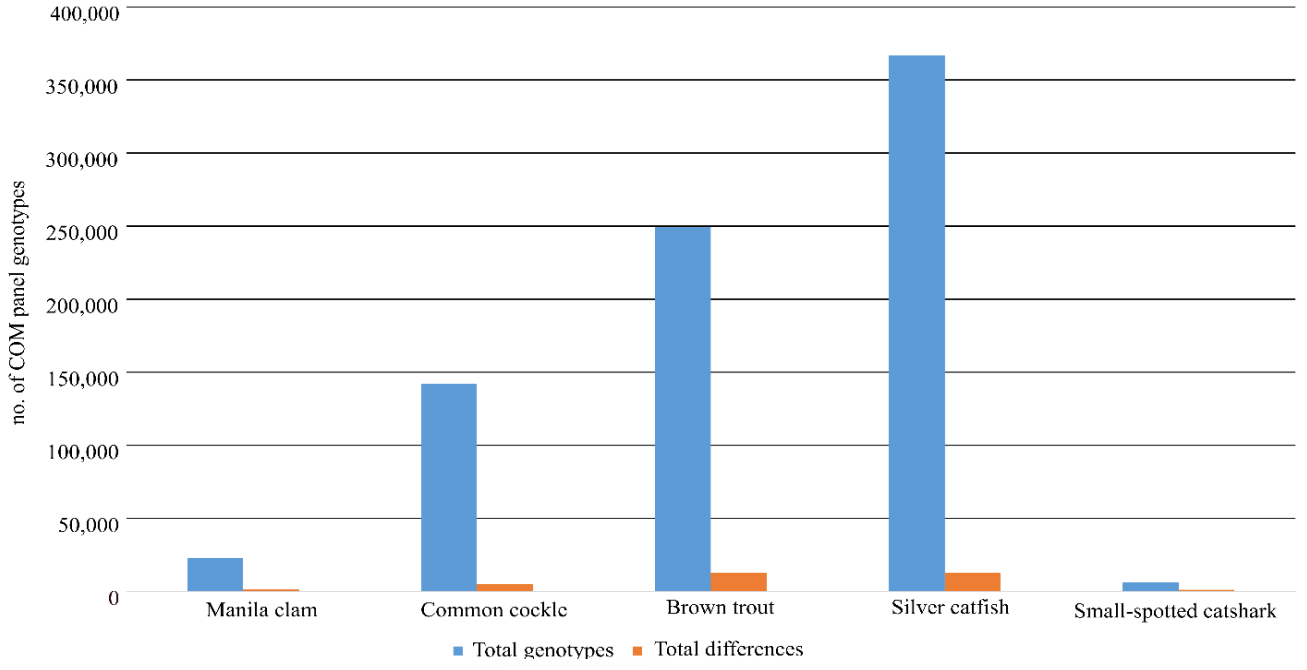

Figure S1. Differences in genotyping between common SNPs (COM panel) from both building-loci pipelines in each species. The total COM genotypes in each species was calculated with this formula: $N_{samples}$ x $N_{COM\ SNPs}$.

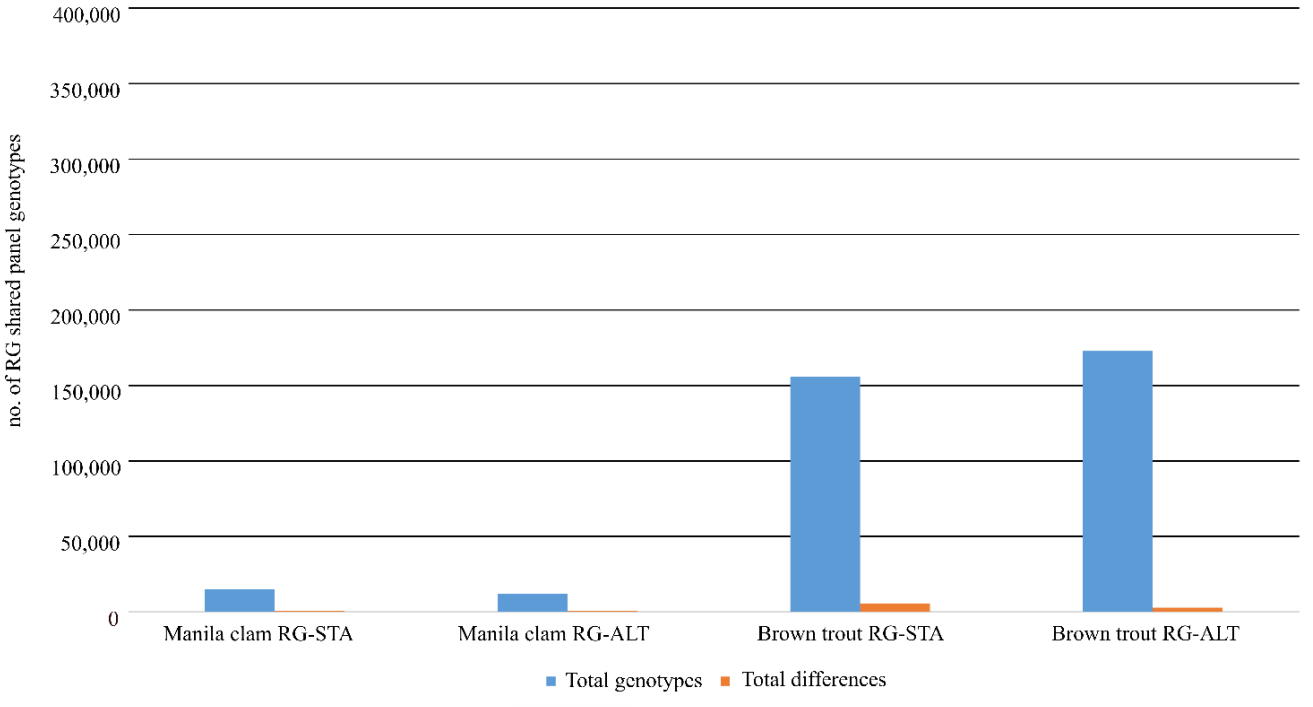

Figure S2. Differences in genotyping between common SNPs from reference genome and *de novo* approach comparisons (i.e. RG-STA and RG-ALT) in Manila clam and brown trout. The total genotypes in each species were calculated with this formula: $N_{samples}$ x N_{SNPs}.

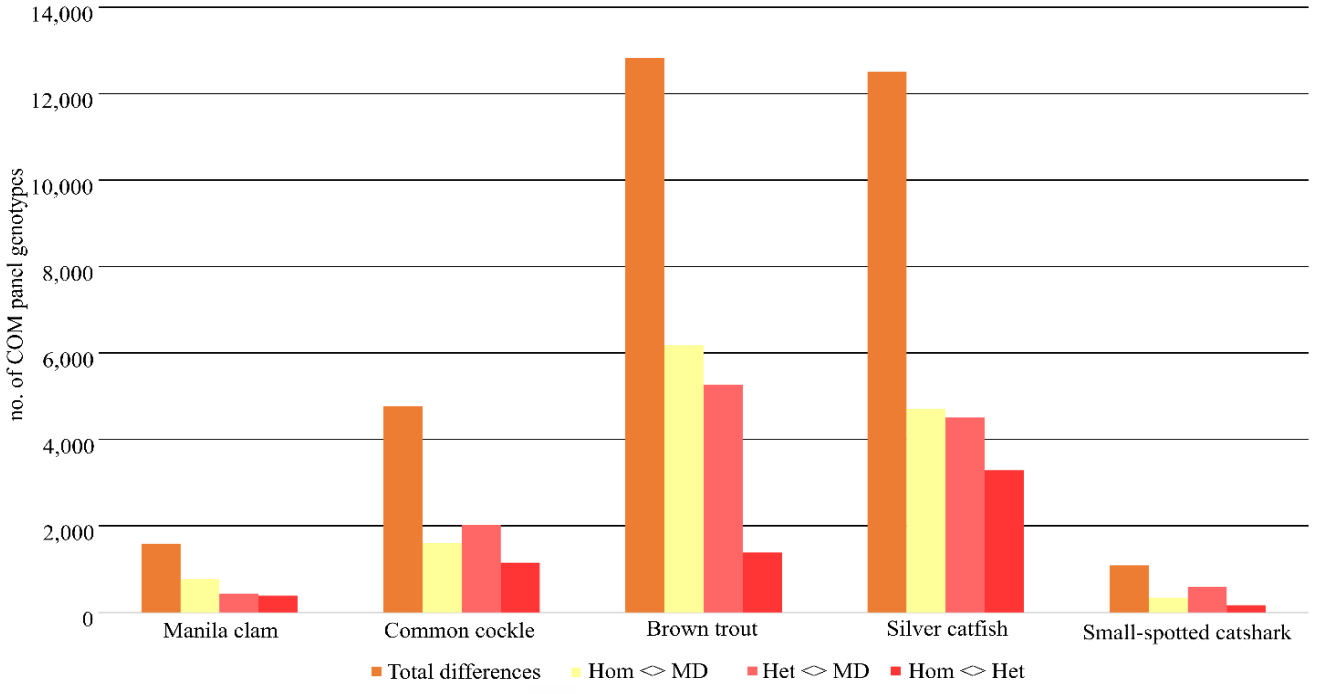

Figure S3. Type of genotype differences between common SNPs (COM panel) from both building-loci **pipelines in each species.** Three types of differences from building-loci pipelines genotyping are represented (i.e. homozygous and missing data, heterozygous and missing data and homozygous and heterozygous). For more detail see Table 2.

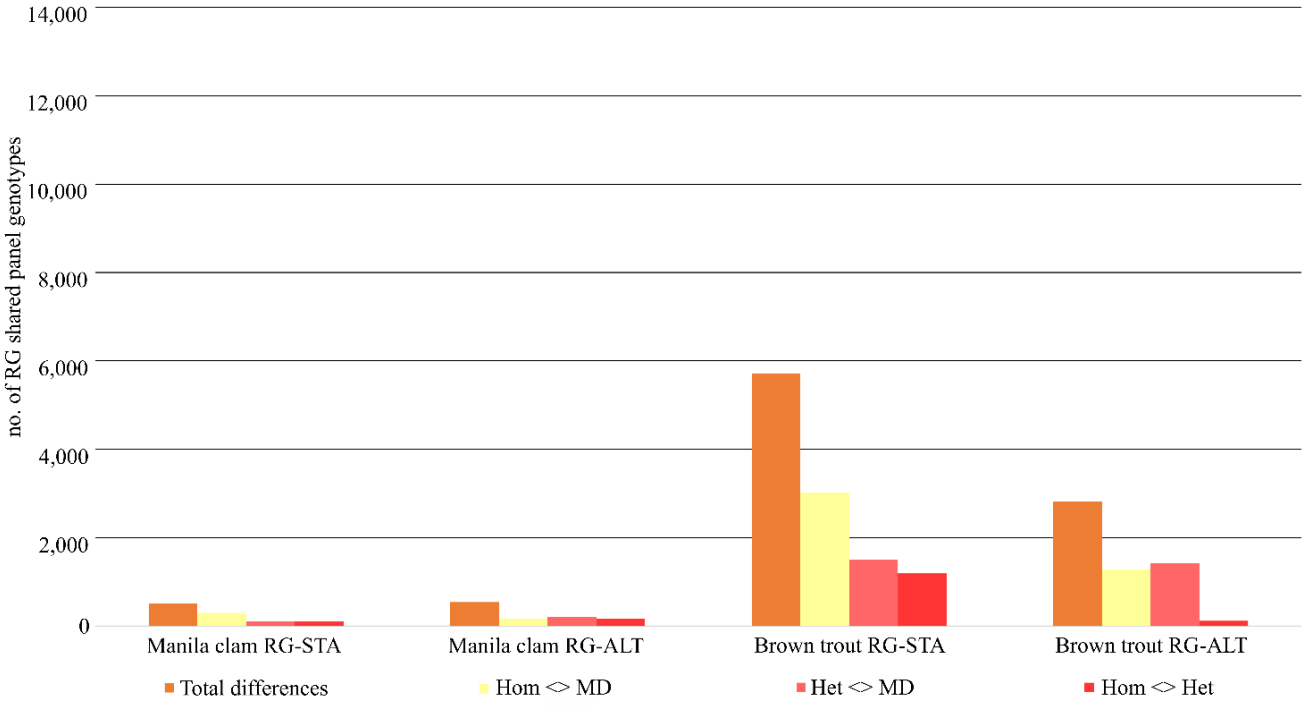

Figure S4. Type of genotype differences between common SNPs from reference genome and *de novo* approach comparisons (i.e. RG-STA and RG-ALT) in Manila clam and brown trout. Three types of differences from building-loci pipelines genotyping are represented (i.e. homozygous and missing data, heterozygous and missing data and homozygous and heterozygous). For more detail see Table 2.

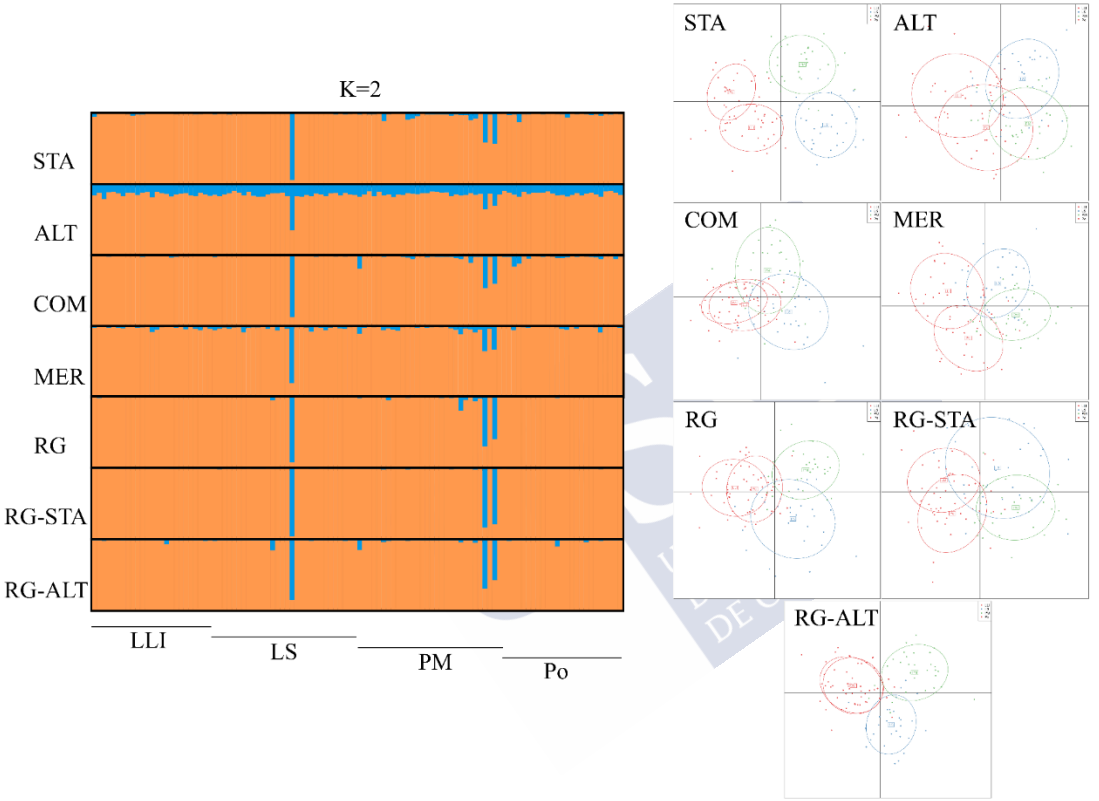

Figure S5. Comparison between CLUMPAK and DAPC outputs for Manila clam (*R. philippinarum*) samples (N = 110). Four locations are included: LLI (Vigo, Spain), LS (Chioggia, Italy), PM (Porto Marghera, Italy) and Po (Po River mouth, Italy).

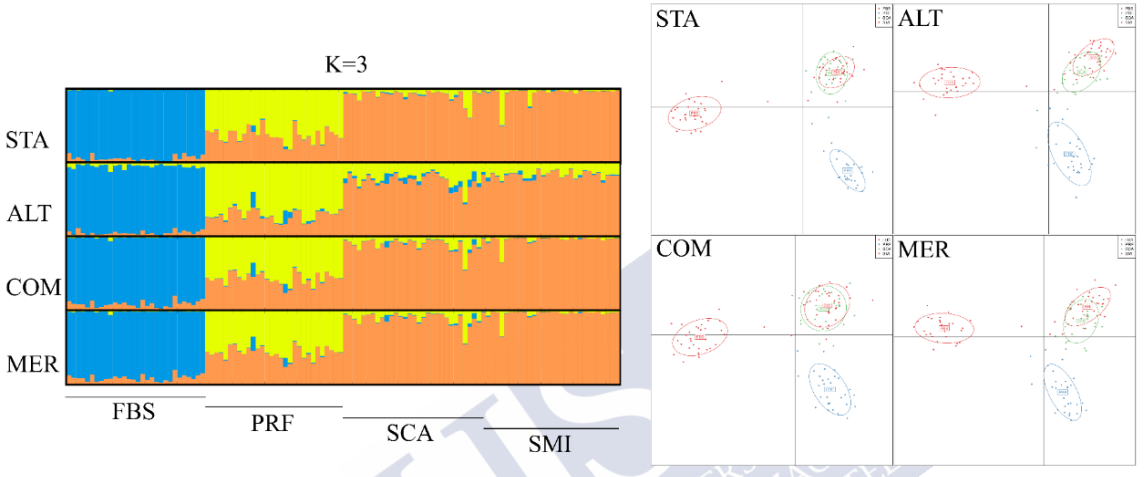

Figure S6. Comparison between CLUMPAK and DAPC outputs for common edible cockle (*C.edule*) samples (N = 120). Four locations are included: FBS (Somme Bay, France), PRF (Ría Formosa, Portugal), SCA (Campelo, Spain), SMI (Miño, Spain).

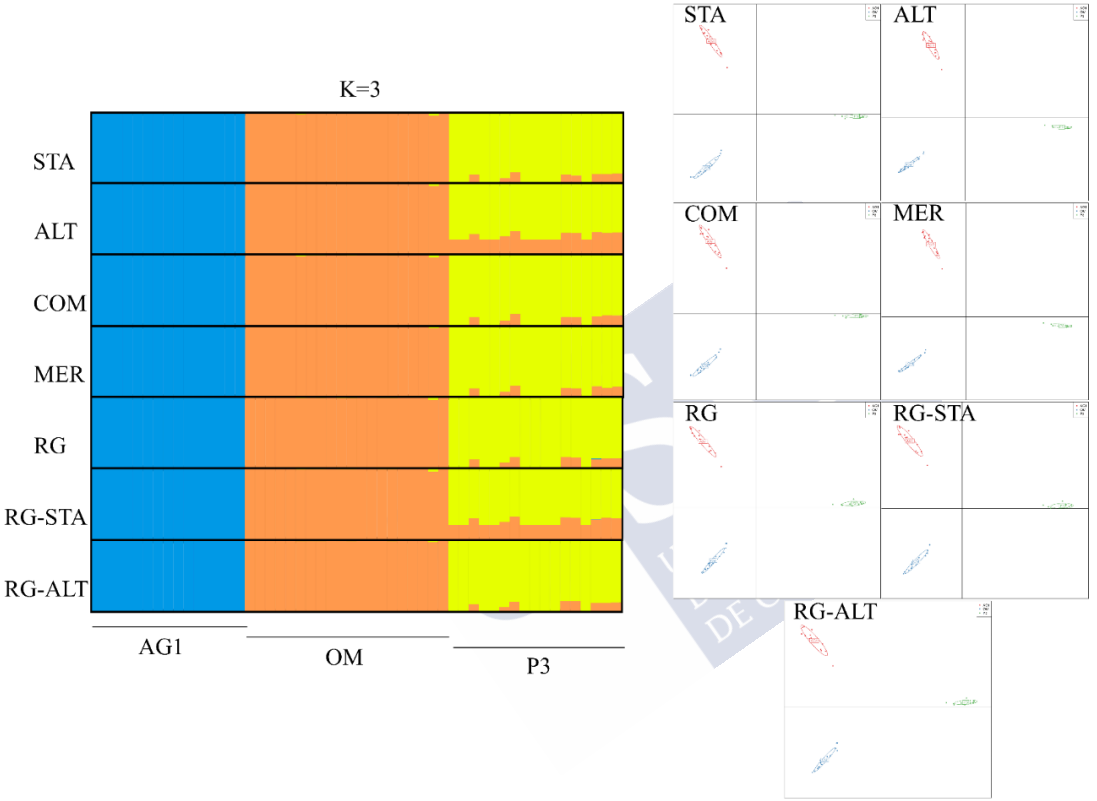

Figure S7. Comparison between CLUMPAK and DAPC outputs for brown trout (*S.trutta*) samples (N = 52). Three locations are included: AG1 (Águeda River, Spain), OM (Omaña River, Spain), P3 (Pisuerga River, Spain).

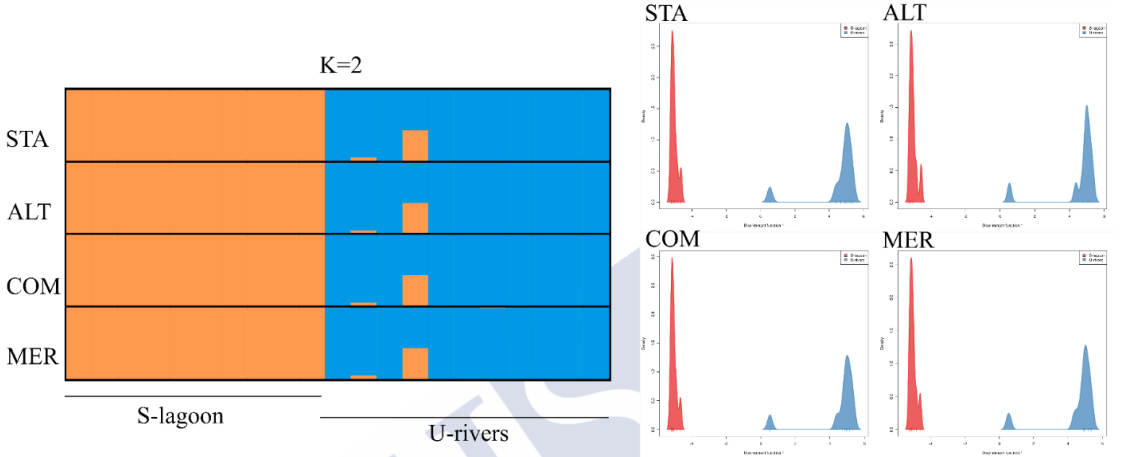

Figure S8. Comparison between CLUMPAK and DAPC outputs for silver catfish (*R. quelen*) samples (N = 21). Two locations are included: S-lagoon (Sauce Lagoon, Uruguay) and U-rivers (Uruguay River basin, Uruguay).

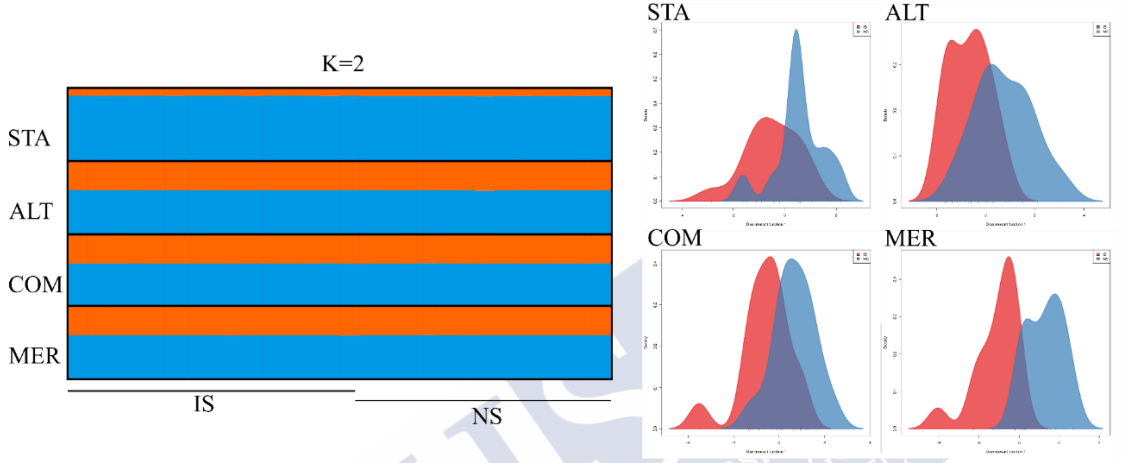

Figure S9. Comparison between CLUMPAK and DAPC outputs for small-spotted catshark (*S. canicula*) samples (N = 28). Two locations are included: IS (Irish Sea), NS (North Sea).